Hot Topics in Astrophysics

2003/2004

Proceedings of the students' workshop held at
Chalmers University of Technology and Göteborg University
6 May 2004

—

Following the course 'Hot Topics in Astrophysics' held at
Chalmers University of Technology and Göteborg University
October 2003 – May 2004

—

Edited by:

Alessandro B. Romeo,

Martin Nord & Markus Janson

Contents

Science with the New Sub-Millimetre APEX Telescope

Christophe Risacher

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(risacher@oso.chalmers.se)

*

Abstract

Sub-millimetre wave astronomy is one of the least explored fields in astronomy. This paper addresses the new science that can be done with it and describes a 12 meter sub-millimetre telescope currently under completion.

1 Introduction

Almost 400 years after Galileo's first telescope, astronomy has developed considerably, covering almost all of the electromagnetic spectrum from Gamma rays, X-rays, Ultraviolet (UV), to Optical, Infrared and Radio astronomy. As the wavelengths are so different, instruments and telescopes techniques vary considerably. Figure 1 shows our Milky Way observed in different bands.

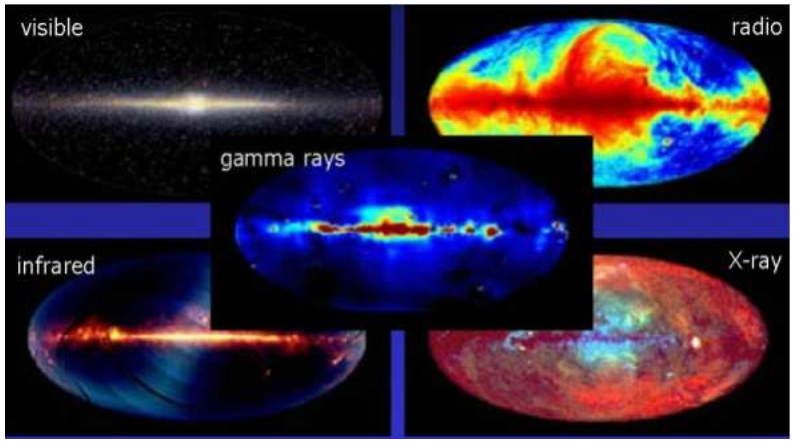

Figure 1: Multi-wavelength astronomy of the milky way sky.

One of the last parts of the electromagnetic spectrum to be studied is sub-millimetre waves. Space is filled with dust, which obscures optical and other wavelengths radiations. Figure 2 shows examples of nebula and star-forming regions full of dust which stops optical light, and as a comparison sub-millimetre radiation is sensitive to the cold dust.

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

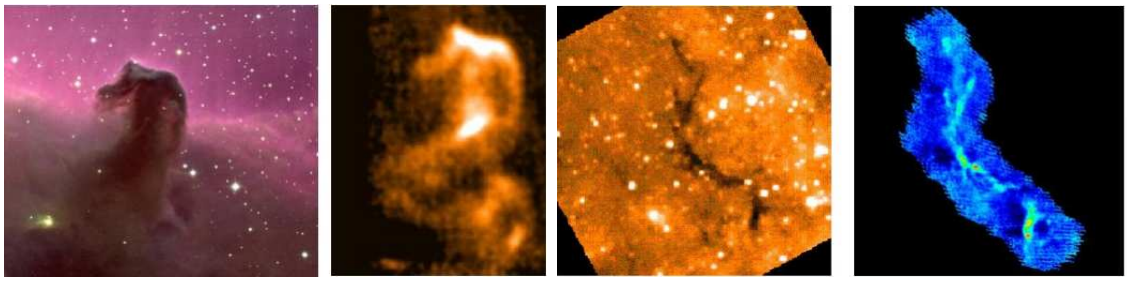

Figure 2: Comparison of optical/sub-millimetre observations of the Horsehead Nebula in Orion (left picture) and in star forming region G11.11-0.12 (right picture).

The first observations in the sub-millimetre band just started about 10 years ago, and very fast development of instruments with ever increasing sensitivity will allow a fast development of that branch of the astronomy. Observations from the Earth are very difficult as its transparency is very low at these wavelengths. Figure 3 shows the atmospheric transmission in different wavelengths and the altitude above sea-level needed to detect the radiations.

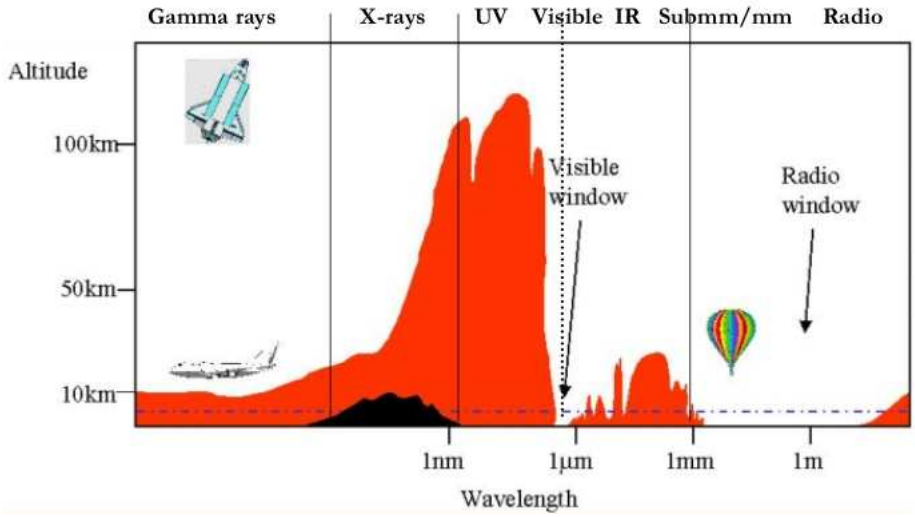

Figure 3: Atmospheric transmission from Earth.

This paper describes a new telescope constructed on one of the best sites on the Earth for astronomy, and discusses what type of science will be done with it.

2 Description of the Site and Telescope

2.1 Site Description - Atmospheric Windows

The selected site for the telescope is Llano Chajnantor, in the Chilean Desert of Atacama, and is at an altitude of 5000 metres. It was selected because it is probably one of the driest places on Earth, and one of the best sites for sub-millimetre wave observations, together with the South Pole and Mauna Kea in Hawaii (Pardo, 2001). Figure 4 shows the selected site, it is a vast plateau currently empty that will host in the near future hundreds of different telescopes. Figure 5 shows the measured atmosphere transmission from 100 GHz to 3 THz, and figure 6 shows the predicted transmission in the wavelength

Figure 4: View of the LLano Chajnantor, a plateau at an altitude of 5 km.

of THz region. Under good weather conditions, observations up to 1.5 THz should be possible from the Earth, something that was previously thought to be possible only from space!

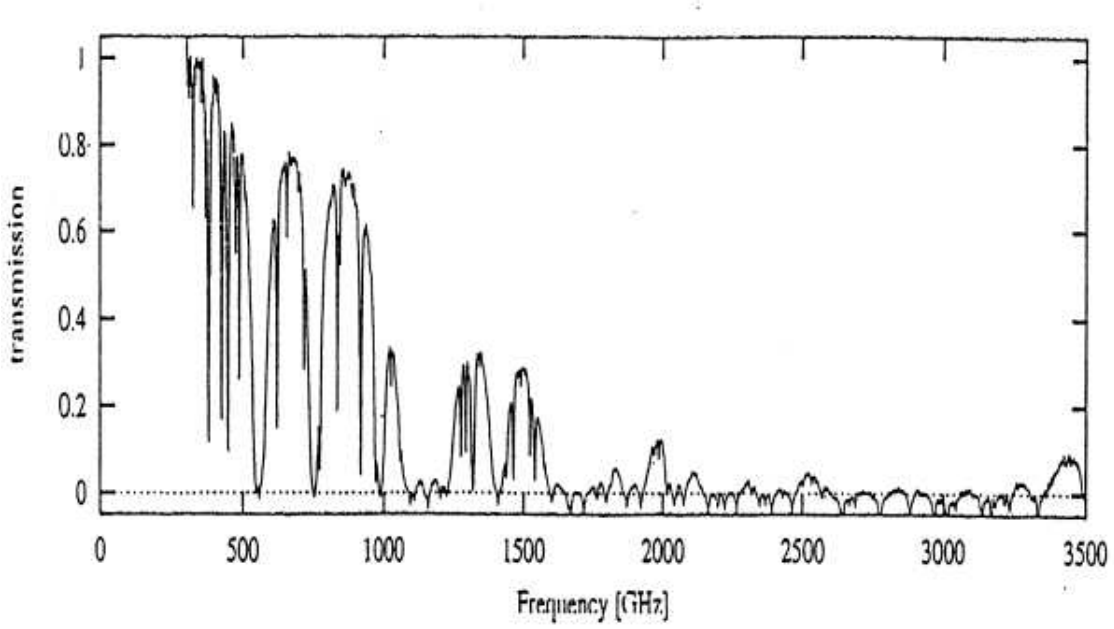

Figure 5: Measured atmospheric transmission at Chajnantor under good conditions (1998 Aug 26). From Paine (2001).

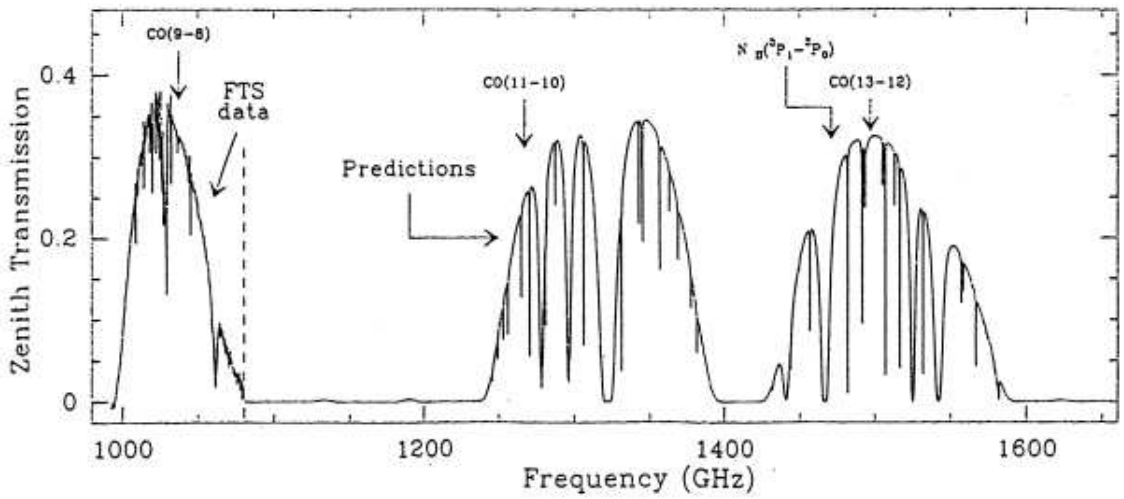

Figure 6: Simulated atmospheric transmission at Chajnantor under good conditions for THz band. From Pardo (2001).

2.2 Telescope Description

The Atacama Pathfinder Experiment (APEX) telescope, is a collaboration between the Max Planck Institute for Radioastronomy (MPIfR)(in collaboration with Astronomisches Institut Ruhr Universitet Bochum (RAIUB), Onsala Space Observatory (OSO) and the European Southern Observatory (ESO) to construct a single dish on the high altitude site of Llano Chajnantor. The telescope was constructed by VERTEX Antennentechnik in Duisburg, Germany. The telescope is currently under installation (Figure 7), and will be fully operational in 2005.

Figure 7: The telescope on the site.

The main antenna characteristics are presented in the table below. Its high surface

accuracy should enable it to observe up to 1.5 THz.

Diameter	12m
Cassegrain/Nasmyth optics	
Mass	125000 kg
f/D	8
Surface accuracy	18 microns
Pointing accuracy	2 arcseconds
Instrumentation	2 Nasmyths and 1 equipment cabin
Frontend	Heterodyne receivers / Bolometer arrays
Backend	Autocorrelators
Coverage	300-1500 microns (230 GHz - 1.2 THz)

The instrumentation to be used on APEX, will include both wide band continuum detectors (bolometers) and heterodyne receivers (single pixel and array cameras) covering the frequency range 150 GHz to 1.5 THz (or 200 μm to 2 mm).

3 What Can We Detect with Sub-Millimetre Astronomy?

The millimetre and sub-millimetre wavebands are especially interesting in that they contain more than 1000 spectral lines of interstellar and circumstellar molecules, atomic fine-structure lines of e.g. carbon, oxygen, and nitrogen as well as continuum from synchrotron radiation from Active Galactic Nuclei (AGN), and thermal emission from cold dust grains (temperatures below 100-200 K)

3.1 Molecular and Atomic Spectral Lines

The most common molecular lines in mm and sub-mm waves are from CO and its isotopes : it traces both high and low density gas. Other molecules like HCN and CS trace high density gas and HCO+ traces ionization. The CO molecule rotational lines seen in emission are excited through collisions with molecular hydrogen therefore it probes dense and warm phases of the molecular gas. This gas component is associated with regions of massive star formation. Higher rotational transitions, requiring temperatures in excess of 100 K and high densities, may be associated with regions close to an Active Galactic Nuclei (AGN). In the mm and sub-mm bands, we can detect not only molecular lines but also the atomic fine-structure lines of carbon, oxygen, and nitrogen. These lines have rest frequencies in the far infrared, but at high-z, they are redshifted into the sub-mm bands. Another advantage of the mm/sub-mm bands is that most molecules have a ladder of spectral lines. If a redshift is so high that a spectral line is shifted out of a given sub-mm window, there is a good chance the next line up the ladder will be shifted into it.

3.2 Thermal Continuum from Dust

Sub-millimetre is most sensitive to cold matter. For example a black-body emission of 10 K cold dust peaks at around 300 μm (Figure 8). Such very cold material is associated with objects in formation, that is the earliest evolutionary stages of galaxies, stars and planets. Figure 9 illustrates the formation of stars and planets.

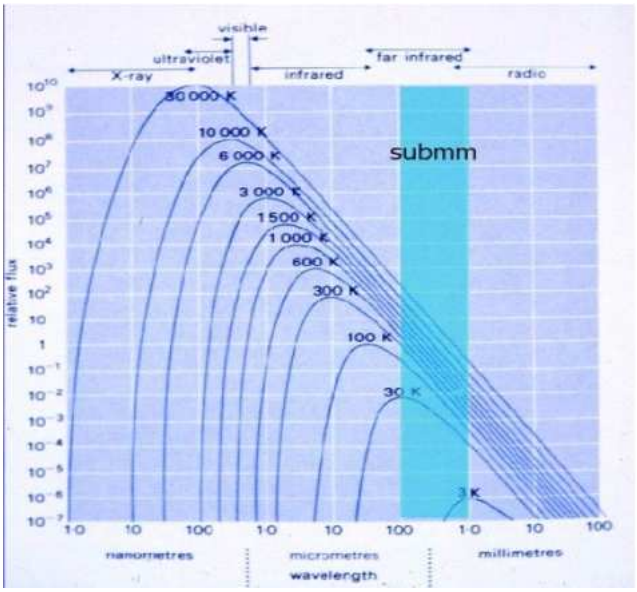

Figure 8: Cold matter ideal black-body peaks in the sub-millimetre regime.

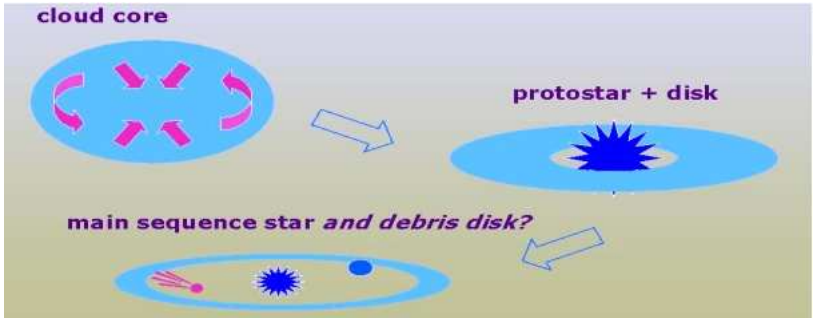

Figure 9: Formation of a planetary system.

Much of the ultraviolet (UV) radiation from massive star formation regions is absorbed by the surrounding dusty clouds. Even the most luminous starburst galaxies are difficult to observe at optical wavelengths. The absorbed radiation is re-emitted by the dust as long-wavelength infrared radiation which can easily escape the star forming regions - but cannot cross the Earth atmosphere. However for very distant objects this radiation is red-shifted to sub-millimeter wavelengths.

Not only is the sub-millimetre the most appropriate regime for cold material, there are other advantages too. Nearly all objects are optically thin, therefore we can see directly into the heart of crucial processes.

3.3 Probing the Very Early Universe

Among the fundamental questions being asked today are: when did galaxies and massive black holes form in the early universe, and how did they subsequently evolve? Modern telescopes are now detecting galaxies out to redshifts beyond 6, thereby probing the "dark ages" where the first stars and black holes may have formed.

By using both spectral lines and continuum , detections of Carbon monoxide (CO) (Figure 10) and dust in the most distant quasar yet known, SDSS J1148+5251 at a redshift of 6.42 were reported in 2003 by Walter (2003). The redshift signifies both distance and epoch and z=6.42 implies that the radiation left the quasar when the Universe was only

about 6% of its current age.

The CO radiation now reveals important information about the density, temperature, and size of the dense molecular center of that galaxy, a region surrounding the massive black hole where a new star is born about every five hours, at a rate much higher than that seen in any galaxy in the local Universe.

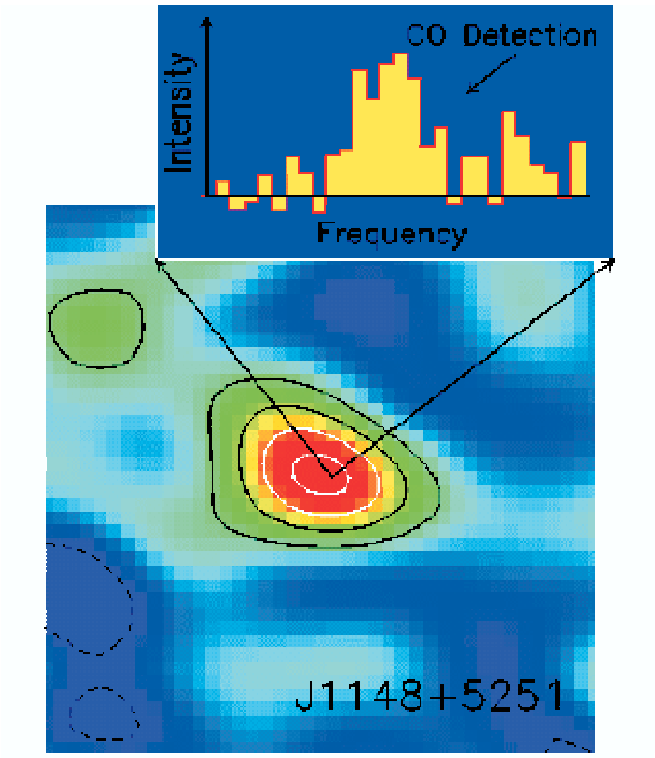

Figure 10: CO detection at z = 6.42. From Bertoldi (2003).

Although dust and carbon monoxide are rare trace components in a gas consisting primarily of hydrogen molecules, the amount of dust and carbon monoxide seen indicates that its formation mechanism must have been very efficient and fast, astronomically speaking.

4 Constraining Cosmological Models

Galaxy clusters are the largest collapsed structures in the Universe. Measuring their distribution and structure provides crucial information on the history and structure of our Universe. Galaxy clusters are embedded in vast amounts of hot, ionized gas - this gas scatters the passing photons of the Cosmic Microwave Background (CMB) and increases their average energy. 1-2% of the CMB photons can be inverse Compton scattered by the hot gas. The resulting distortion in the CMB is called the Sunyaev-Zel'dovich (SZ) Effect and can be used as a sensitive probe of cosmological models and cluster physics. The bolometer arrays at APEX will have an ideal spatial resolution and sensitivity to measure the SZ effect toward distant clusters - the large 2 mm bolometer array planned for APEX will perform the first large-area survey of distant galaxy clusters, for example.

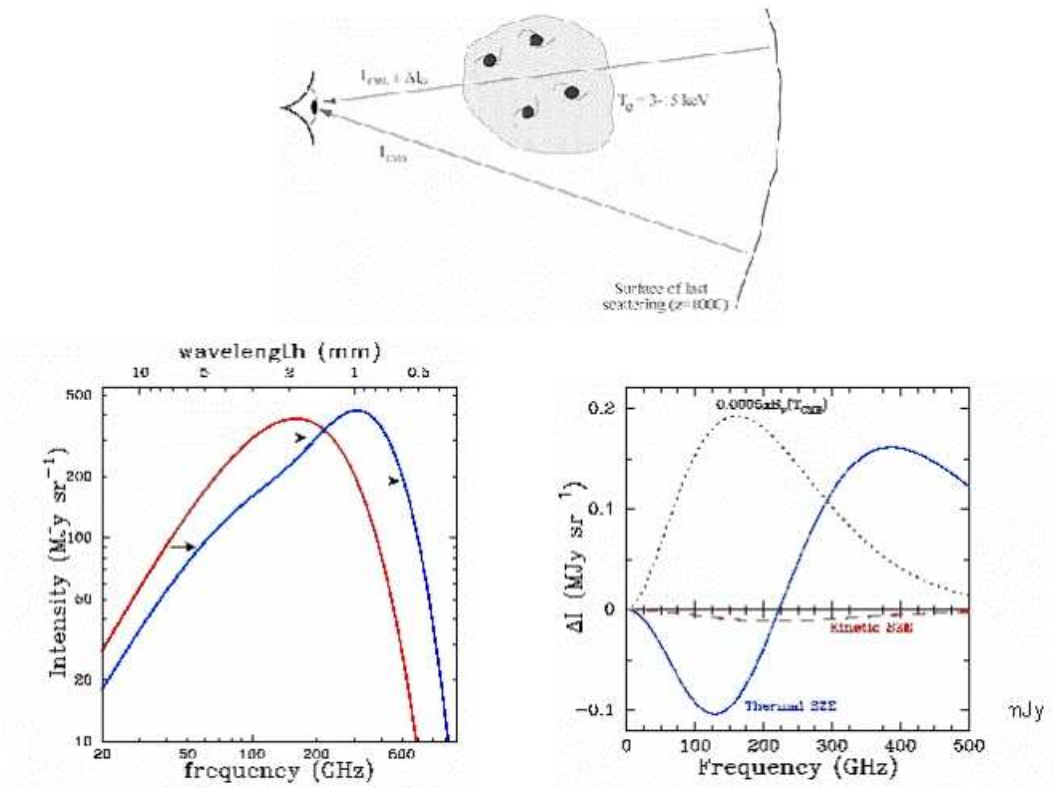

Figure 11: Top drawing illustrates the ionized gas distorting the cosmic microwave background. In the left picture, the ideal back-body of the cosmic microwave background (red) is compared to the disturbed one by the SZ effect (blue). Right picture shows the thermal and kinetic SZ effects.

Deviations in intensity from the black-body ideal of the CMB are shown, as a function of frequency, as the blue line in Figure 12.

There is a second SZ effect, the kinetic SZ effect, caused by the motion of of the galaxy clusters with respect to the rest frame of the CMB. The net motion of the scattering electrons in the hot intercluster gas imparts a Doppler shift to the scattered photons. This kinetic SZ effect is shown in the graph as the red dotted line.

Among other things, observing with APEX will allow to discover and catalog 1000 previously unknown galaxy clusters in a mass limited survey, it will constrain the mass density of the Universe Ωm and dark energy equation of state ω, measure the Hubble constant H_0 and acceleration parameter q_0 independent of the distance ladder.

5 Southern Hemisphere Sky

APEX will be able to completely map unique objects at sub-millimeter wavelengths. Some of the most interesting sources in the sky can best (or only) be studied from the southern hemisphere, including the Galactic center (an important prerequisite for understanding the central regions of other galaxies), the Magellanic clouds (the nearest galaxies to our own and prototypes of metal-poor galaxies in an earlier stage of evolution), and Centaurus A (the nearest galaxy with an active nucleus).

5.1 The Galactic Center

The Galactic Center (Figure 12) provides an opportunity to test current ideas about star formation, activity, and stellar dynamics in the nuclei of galaxies. The fact that it transits almost overhead at Chajnantor is of great advantage. Since it shares many properties of other, more spectacular galactic nuclei, a detailed investigation of the Galactic center and its surroundings is a necessary step to gain a better understanding of the physical processes governing galactic nuclei in general.

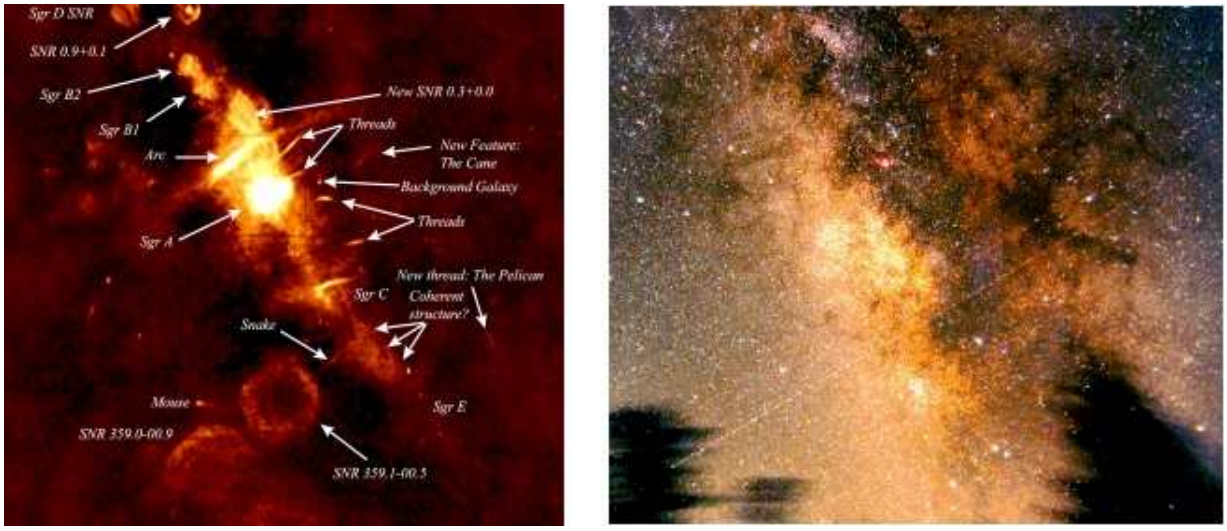

Figure 12: Radio view of the galactic center (left) compared to optical view (right).

5.2 The Magellanic Clouds

The Magellanic Clouds are excellent laboratories to study galaxy evolution. For instance, they are metal deficient in comparison to the Milky Way and star formation in these conditions should be more similar to galaxies in the distant universe.

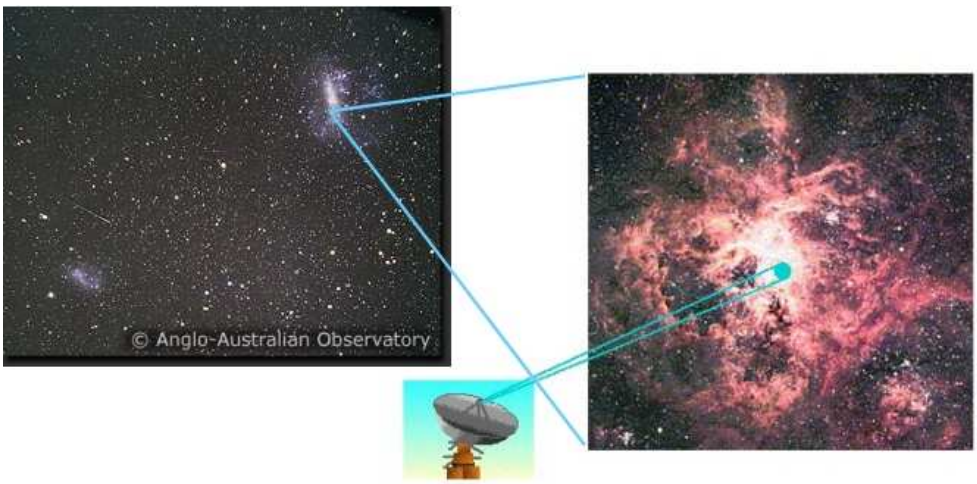

Figure 13: 30 DORADUS (NGC 2070 = Tarantula Nebulae) is a star-forming region of the Large Magellanic Cloud. The APEX beam will be able to resolve regions of 2-7 parsec in size. The blue circle in the picture shows the size of the largest beam of APEX.

The proximity of the local group and the southern location of the magellanic clouds make them excellent targets for APEX which will be able to resolve star-forming regions such as the 30 Doradus in the Large Magellanic Cloud (Figure 13).

The Local Group is also a very good laboratory for studying galaxy interaction. For instance, the Milky Way is in the process of accreting a small galaxy (the Sagittarius Dwarf galaxy) and is in close contact with the Magellanic Clouds via a stream of gas called the Magellanic Stream.

5.3 Centaurus A

The nearby (about 3 Mpc away) giant elliptical galaxy NGC 5158 hosts the strong radio source Centaurus A, with radio lobes extending out to 2 - 3 degrees on each side of the nucleus (Figure 14). Embedded in a massive warped molecular disk is a faster rotating circumnuclear ring at radius about 100 pc, oriented perpendicular to the inner radio jet. A variety of molecular tracers have been observed in absorption against the strong Centaurus A radio continuum at millimeter wavelengths. At sub-millimeter wavelengths it is possible to further elucidate the physical properties by means of spectroscopy of atomic (e.g., carbon) and excited molecular species, and studies of the dust continuum in this edge-on system.

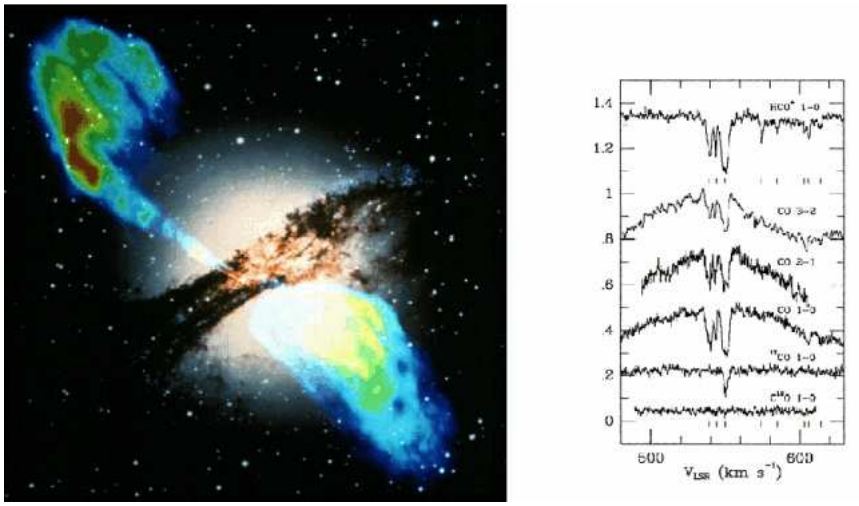

Figure 14: Optical picture of Cen A, with the radio jet overlaid. Toward the center, CO and HCO+ spectra, showing broad nuclear emission and narrow foreground absorption, are observed.

5.4 Summary of the Science with APEX

APEX will allow astronomers to study warm and cold dust in star-forming regions both in our own Milky Way and in distant galaxies in the young universe. High frequency spectral lines enable the exploration of the structure and chemistry of planetary atmospheres, evolved stars, molecular clouds as well as inner regions of starburst galaxies. It will cover a wide range of studies from the vast scales of the structure of the Universe down to the physics and chemistry of comets.

6 APEX, Pathfinder of ALMA Project

APEX will also serve as a pathfinder for the Atacama Large MIllimetre Array (ALMA), a joint US/European collaboration, in all of its wavelength ranges. It will carry out wide-field observations, identifying and compiling potential sources for a later detailed study with ALMA, thus allowing a highly efficient use of ALMA observation time when ALMA goes operational.

6.1 ALMA project

ALMA will be the largest ground based astronomy project of the next decade after the optical projects VLT/VLTI and together with the Next Generation Space Telescope (NGST), one of the two major new facilities for world astronomy possibly coming into operation by the end of the decade. ALMA will be comprised of 64 12-meter sub-millimetre quality antennas (Figure 15), with baselines extending up to 10 km. Its receivers will cover the range from 30 to 950 GHz. Baselines from 20 m to 14 km. An example of typical angular resolution is 0.05 arcseconds at 100 GHz.

Figure 15: Artistic impression of the future array of 64 antennae of 12m diameter on the Chajnantor plateau.

This interferometer will allow to reach an unprecedented sensitivity and resolution at millimetre and sub-millimetre wavelengths As examples of scientific goals, il will allow to study in much deeper details selected objects found by pathfinder telescopes as APEX. It will also study protoplanetary disks where planets are about to form, planetary atmospheres and surfaces, comets and minor bodies in the Solar System, evolved stars circumstellar envelopes, winds... It will measure physical properties and molecular abundances in dense clouds, study merging galaxies, or gamma-ray bursters, etc.

7 Conclusions

Sub-millimetre astronomy is a largely unexplored part of the electromagnetic spectrum. It is now in full development. New telescopes that are being built or will be like APEX and ALMA will help to understand better many interesting problems in astrophysics, like how galaxies, stars and planet formed or what the final fate of the universe is.

Acknowledgements

I would like to thank Alessandro Romeo and all the students for their enthusiasm and help during this course.

References

Alma science page `http://www.eso.org/projects/alma/science/`

Apex home page Onsala `http://www.oso.chalmers.se/oso/apex/index.html`

Apex home page MPIfR `http://www.mpifr-bonn.mpg.de/div/mm/apex.html`

Bertoldi F., et al. 2003, A&A 409, 47

Paine S., Blundell R., Papa D.C., Barrett J.W., Radford S.J.E. 2000, PASP, 112, 108

Pardo J.R., Serabyn E.,Cernicharo J. 2001, JQSRT, 68, 419

Walter F., Carilli C., Bertoldi F., et al. 2003, Nature 424.

Wavelets in Radio-Astronomy

Rajat Mani Thomas

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(rajat@etek.chalmers.se)

*

Abstract

Due to the poor sampling of the U-V plane (or the Fourier space), which is a consequence of finite number of interferometres, image retreival (or deconvolution) in Radio Astronomy has been a long standing problem. Conventional method like CLEAN seem to diverge in many cases, especially when the sources are extended. We have tried to modify the CLEAN by introducing multiresolution analysis. The idea behind it is that the sky appears empty at some given resolution, whatever the source structure. This produces a convergent CLEAN at that resolution. Thus by decomposing the dirty-map and dirty-beam and applying CLEAN at each level, deconvolution can be achieved in a better and faster way. We study the conventional CLEAN and implement *Multi-Resolution CLEAN* (MRC) on the test-images and also on images from the Mauritius radio-telescope and then make a comparative study. The results obtained were encouraging and show an improvement in obtaining better deconvolved images.

1 Introduction

Unlike in optical astronomy images captured by Radio-telescopes are limited in their quality due to the very low signal to noise ratio. In order to extract useful information from radio sources, several stages of processing of the acquired data is needed. The stages, in brief, involve averaging out of the noise, specialized De-noising techniques and Deconvolution (see e.g., Chapter-1 Kraus John D, 1986, Burke & Smith 2002),. Our task involved the study and application of a recently proposed technique for deconvolution based on multi-resolution approach. The concepts and the mathematical formalisms in wavelet theory and multi-resolution analysis (MRA) enabled us understand and implement the proposed algorithm.

1.1 Statement of the Problem

The signal or data from the cosmos is captured by a Radio Telescope and is then processed. While capturing the data, the incoming signal gets convolved with the Radio Telescope's

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

antenna beam pattern, also known as the Point Spread Function or PSF of the telescope. For a uniformly illuminated aperture the PSF will be a product of sinc functions along the declination and right ascension (Ref. Glossary). Now if we consider radio point-sources in the sky, they can be thought of as representing Dirac-delta functions in space. The convolution process thus causes the repetition of this PSF at every source location, giving rise to unwanted sidelobes for a source. If the two radio-sources are relatively close to each other, then the sidelobes of these sources will interfere with each other. If the interference in constructive then a false source may show up on the map.

The process of eliminating the effect of the PSF of an antenna and retrieving the information about sources in the sky is the problem of Deconvolution. Most of the earlier techniques failed to achieve this to the extent required when the sources are extended or are extremely close to each other. Our aim is to implement this technique called the Multi-resolution CLEAN (see starck & Bijaoui 1994, Bhat, Cordes & Chatterjee 2003,).This method is based on wavelet transform of the Images.

1.2 Radio Interferometry and Aperture Synthesis

The angular resolution, or ability of a Radio telescope to distinguish fine details in the sky, depends on the wavelength of observation divided by the size of the instrument, $\theta = (1.22\lambda/d)$. Yet, even the largest antennas, when used at their shortest operating wavelength, have an angular resolution only a little better than one arc minute, which is comparable to that of the unaided human eye at optical wavelengths. Because radio telescopes operate at much longer wavelengths than do optical telescopes they must be much larger than optical telescopes to achieve the same angular resolution.

At radio wavelengths, the distortions introduced by the atmosphere are less important than at optical wavelengths, so the theoretical angular resolution of a radio telescope can in practice be achieved even for the largest dimensions. High angular resolution at radio wavelengths is achieved by using interferometry to synthesize a very large effective aperture from a number of small elements. In a simple two-element radio interferometer, the signals from an unresolved, or point-source alternately arrive in phase and out of phase as the earth rotates and causes a change in the difference in path from the radio source to the two elements of the interferometer (see eg., Wakker & Schwarz 1991). This produces interference fringes in a manner similar to that in an optical interferometer. If the radio source has finite angular size, then the difference in path length to the elements of the interferometer varies across the source. The measured interference fringes from each interferometer pair thus depend on the detailed nature of the radio "brightness" distribution in the sky. Each interferometer pair measures one "Fourier component" of the brightness distribution of the Radio Source (see Burke and Smith 2002 for a detailed treatment of Interferometry).

Work by Australian and British radio astronomers in the 1950s and 1960s showed that movable antenna elements combined with the rotation of the earth can sample a sufficient number of Fourier components with which to synthesize the effect of a large aperture and thereby reconstruct high-resolution images of the Radio-Sky. In recognition of their contributions to the development of the Fourier synthesis technique, more commonly known as aperture synthesis, or earth-rotation synthesis, Martin Ryle and Antony Hewish were awarded the 1974 Nobel Prize for Physics. During the 1960s the Swedish Radio Astronomer, Jan Högbom developed a technique called "CLEAN," which is used to remove the spurious responses from a celestial radio-image caused by the use of discrete, rather than continuous spacing in deriving the radio-image.

1.3 Convolution - a Physical Interpretation

Our daily life abounds with phenomena that can be described mathematically by convolution. Spreading, blurring, and mixing are qualitative terms frequently used to describe these phenomena. Sometimes the spreading is caused by physical occurrences unrelated to our mechanisms of perception; sometimes our sensory inputs are directly involved. The blurred visual image is an example that comes to mind. The blur may exist in the image that the eye views, or it may result from a physiological defect. Biological sensory perception has parallels in the technology of instrumentation. Like the human eye, most instruments cannot discern the finest detail. Instruments are frequently designed to determine some observable quantity while an independent parameter is varied. An otherwise isolated measurement is often corrupted by undesired contributions that should rightfully have been confined to neighbouring measurements. When such contributions add up linearly in a certain way, the distortion may be described by the mathematics of convolution.

These spreading and blurring phenomena profoundly affect radio-stronomy. The recovery of an image as it would be observed by a hypothetical, perfectly resolving instrument is an exciting goal. Recent developments have stimulated development of restoring methods that receive increasingly wider applications. Whether we concern ourselves with the science that requires imaging or other experimental approaches, our state of knowledge is often defined by the resolving limit of our instruments. Through modern restoring methods, the scientist has access to information that would otherwise remain unavailable.

Convolution in its simplest form is usually considered to be a blurring or smoothing operation. Typically, each output value is identified with a corresponding input value. It is obtained, however, by processing that value and some of its neighbours. One simple way of smoothing the fluctuation is to perform a moving average (Jansson 1997).

The behavior of a linear, continuous-time, time-invariant system with input signal $\mathbf{x(t)}$ and output signal $\mathbf{y(t)}$ is described by the convolution integral (1)

$$y(t) = \int_{-\infty}^{\infty} x(v)h(t-v)dv \tag{1}$$

The signal $\mathbf{h(t)}$, assumed known, is the response of the system to a unit impulse input. To compute the output y(t) at a specified t, first the integrand $\mathbf{h(v)\ x(t\text{-}v)}$ computed as a function of $\mathbf{v}$. Then integration with respect to $\mathbf{v}$ is performed, resulting in $\mathbf{y(t)}$. And similarly in the discrete domain convolution can be expressed as (2);

$$y(n) = \sum_{m=-\infty}^{\infty} x(m)h(n-m) \tag{2}$$

Here $\mathbf{h(n)}$ is the response of the system, $\mathbf{x(n)}$ the input and $\mathbf{y(n)}$ the output.

2 Deconvolution - the Formalism

In the previous section we developed the concept of the convolution integral and its discrete version. Even considering the brief treatment of this topic, we may have enough information to foresee the difficulties that could arise when we attempt to compute values of $\mathbf{x(n)}$ from given values of $\mathbf{y(n)}$ and $\mathbf{h(n)}$ in (2).

Such difficulties do indeed occur. The problem of deconvolution has therefore been the subject of vast literature spanning the numerous special fields of science. Deconvolution

is a key area in signal and image processing. It is used in a number of areas like, image de-blurring, correction of spherical aberration in mirrors, image sharpening and image restoration in radio-astronomy. Our Project is related to the *"Deconvolution of Radio Images"*.

Let us begin the formalism of the deconvolution problem by reformulating it in the continuous domain as(3),

$$i(x) = \int_{-\infty}^{\infty} o(x\prime)s(x - x\prime)dx\prime \tag{3}$$

In astronomy, we seek the object function $\mathbf{o(x)}$ when we are given the image data $\mathbf{i(x)}$ and the known point spread function $\mathbf{s(x)}$ of the intervening radio telescope. The point-spread function or the PSF of a Radio Telescope is a measure of how it degrades a point source. The name is based on the fact that in optical phenomenon, the impulse corresponds to a point of light and that an optical system responds by blurring (spreading) the point, with the degree of blurring being determined by the quality of optical components. In the radio astronomy adaptation, the object $\mathbf{o}$ is usually the true image of interest, as it would be observed by an ideal, perfectly resolving telescope.

Having described the varies quantities above, a deconvolution problem can be stated as: *"If* $\boldsymbol{i}$ *represents the known data,* $\boldsymbol{s}$ *the blurring phenomena, the effects which are to be eliminated, then* $\boldsymbol{o}$ *is the function that we seek, free of the spreading due to* $\boldsymbol{s}$*"*.

If we are observing an image $\mathbf{o(x\prime)}$ with the aid of an instrument having a characteristic response function $\mathbf{s(x\text{-}x\prime)}$, then $\mathbf{i}$ represents the data acquired. If we have a perfectly resolving instrument, then $\mathbf{s(x\text{-}x\prime)}$ is a Dirac delta function, and our data i(x) directly represent the true object, that is,$\mathbf{o(x)}$. In this case we have no need for deconvolution

Conversely, we may observe a point source such that the output $\mathbf{o(x\prime)}$ is approximated to δ(x$\prime$). Now the data $\mathbf{i(x)}$ represents the response function. In some fields, such as process control, determining the spread function is a special case of a more general technique called system identification.

Given the difficulty of the deconvolution problem, it may seem outrageous to suggest that simultaneous extraction of both object and point spread function is possible. This feat, however, which is called blind deconvolution (see Siddhichai & Chambers 2002), has been accomplished in some specialized applications with increasing success. Many methods that have been developed for deconvolution rely on the continuous functional representation of $\mathbf{s(x\text{-}x\prime)}$ and $\mathbf{i(x)}$. These methods are limited in usefulness for the experimentalists, who wish to apply deconvolution techniques by computer to digitized images or data.

3 Deconvolution in Radio Astronomy - the Method CLEAN

3.1 The Högbom CLEAN

The standard CLEAN or the Hgbom CLEAN (see Högbom 1974) relies on a qualitative constraint: it assumes that the sky brightness is essentially an ensemble of point sources (the sky is dark, but full of stars). The algorithm, which derives from such an assumption, is straightforward. It is a simple "matching pursuit"

1. Initialize a Residual map to the Dirty map
2. Initialize a Clean component list to zero.

3. Assume strongest feature in Residual map originates from a point source

4. Add a fraction γ (the Loop Gain) of this point source to the Clean component list, remove the same fraction, convolved with the dirty beam, from the Residual map.

5. If the strongest feature in the Residual map is larger than some threshold, go back to point 3 (each such step is called an iteration).

6. If the strongest feature is below threshold, of if the number of iterations N_{iter} is too large, go to point 7.

7. Convolve the Clean component list by a properly chosen Clean Beam (this is called the restoration step).

8. Add to the result the Residual map to obtain the Clean Map.

The CLEAN algorithm has a number of free parameters. The loop gain controls the convergence of the method. In theory, $0< \gamma <2$, but in practice one should use $\gamma \approx 0.1$ - 0.2, depending on sidelobe levels, source structure and dynamic range. While high values of γ would in principle give faster convergence, since the remaining flux is $\propto (1-\gamma)^{N_{iter}}$ if the object is made of a single point source, deviations from an ideal convolution equation force to use significantly lower values in order to avoid nonlinear amplifications of errors. Such deviations from the ideal convolution equation are unavoidable because of thermal noise, and also of phase and amplitude errors, which distort the dirty beam.

The threshold for convergence and number of iterations define to which accuracy the deconvolution proceeds. It is common practice to CLEAN down to about the noise level or slightly below. However, in case of strong sources, the residuals may be dominated by dynamic range limitations.

The clean-beam used in the restoration step plays an important role. It is usually selected as a 2-D Gaussian, which allows the convolution to be computed by a simple Fourier transform, although other choices could be possible. The size of the clean beam is a key parameter. It should be selected to match the (inner part of) dirty beam; otherwise the flux density estimates may be incorrect. An example is a dirty beam with narrow central peak on top of a broad "shoulder". Small-scale structures will be properly reconstructed, but larger ones not. The last step in the CLEAN method plays a double role. On one hand, it protects against insufficient deconvolution. Furthermore, since the residual image should be essentially noise if the deconvolution has converged, it allows noise estimate on the cleaned image.

3.2 The Limitations of CLEAN

Non-uniqueness: A major drawback to the use of 'CLEAN' is the way in which its answers depend upon the various control parameters: the location of 'CLEAN' boxes, the loop gain and the number of 'CLEAN' subtractions. By changing these one can, even for a relatively well-sampled u,v plane, produce noticeably different final images. In the absence of an error analysis of 'CLEAN', one can do nothing about this except practice vigilance and avoid interpreting any aspects of an image that are unstable to the choice of control parameters. Part of our purpose is to illustrate the effects that should keep you from believing completely in the final images produced by 'CLEAN'. In almost any astronomical application, Monte Carlo tests of 'CLEAN', and comparisons of its results

with those of other deconvolution methods, are illuminating. They remain the only practical way to estimate the effects of data errors and of different 'CLEAN'ing strategies on the final image. Eventually, we will gain experience of applying 'CLEAN' to a wide range of different images. This experience will let you guide 'CLEAN' to plausible results more quickly. The 'CLEAN' images that you then produce may not be intrinsically more reliable, but you will have calibrated your use of them for astrophysics much better!

Theoretical understanding of 'CLEAN' is relatively poor even though the original algorithm is quite old. Schwarz analyzed the Hgbom 'CLEAN' algorithm in detail (see Schwarz 1978) . He notes that in the noise-free case the least-squares minimization of the difference between observed and model visibility, which 'CLEAN' performs, produces a unique answer if the number of cells in the model is not greater than the number of independent visibility measurements contributing to the dirty image and beam, counting real and imaginary parts separately. This rule is unaffected by the distribution of u,v data so that, in principle, super-resolution is possible if enough visibility samples are available. In practice, however, the presence of noise and the use of the FFT algorithm to calculate the dirty image and beam corrupt our knowledge of the derivatives of the visibility function upon which super-resolution is based. Clearly, even if the FFT is not used, the presence of noise means that independence of the data must be redefined. Schwarz produced a noise analysis of the least-squares approach but it involves the inversion of a matrix of side NlNm and so is impractical for large images; furthermore, we are really interested in 'CLEAN', not the more limited least-squares method, since 'CLEAN' will still produce a unique answer in circumstances where the least-squares method is guaranteed to fail. To date no one has succeeded in producing a noise analysis of 'CLEAN' itself. The existence of instabilities in 'CLEAN' makes such an analysis highly desirable.

Instability: One instability of 'CLEAN' is well known: its images of extended sources are sometimes modulated at spatial frequencies corresponding to un-sampled parts of the u,v plane (Cornwell 1983). Convolution with a larger 'CLEAN' beam than usual can mask this problem, especially if the un-sampled regions are in the outer parts of the u,v plane. Reducing the loop gain to very low values generally has little effect. Various modifications to CLEAN have been invented to try to combat this problem, but overall the experience is that the best solution is to use another deconvolution algorithm, such as MEM or Maximum Entropy Method (see Narayan & Nityananda 1986).

The occurrence of the stripes is a natural consequence of the incorrect information about Radio Sources embodied in the 'CLEAN' algorithm. Astronomers have not found much evidence for real stripes in Radio Sources, so they are skeptical about stripes in 'CLEAN' images. Unfortunately the only a priori information built into 'CLEAN' is that astronomers prefer to see mainly blank images; there is no bias against stripes. These and other considerations motivated the development of deconvolution algorithms which incorporate extra constraints on astrophysically plausible brightness distributions or which are claimed to produce, in some way, optimal solutions to the deconvolution equation.

4 Multi-Resolution Analysis (MRA) -Wavelets

A wavelet is a small wave with its average value being zero and its energy concentrated in time. This provides a method for analyzing transient, non-stationary or time varying

phenomenon. The basis functions for wavelet analysis are the translated and dilated versions of what is called a mother wavelet. A mother wavelet Ψ(t) has the characteristics,

$$\left.\begin{array}{c} \int_{-\infty}^{\infty} \psi(t)dt = 0 \\ \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \end{array}\right\} \tag{4}$$

Unlike Fourier analysis, a wavelet expansion is not a unique representation of the function. We can have different mother wavelets and its corresponding basis sets. For a given function f(t) one can define both Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). CWT is mainly used for theoretical purposes. It has a lot of redundancy. DWT is easy to implement and it provides a way for multi-resolution analysis (MRA).

The central result used in DWT is that every square integrable function f(t), i.e., f(t)$\epsilon L^2(R)$,can be represented by a linear combination of the dyadic (powers of 2) dilates and translates of wavelet Ψ(t) as,

$$f(t) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a(j,k)\Psi(2^{-j}t - k) \tag{5}$$

Where a(j,k) is the Discrete Wavelet Transform coefficient at translation k at a resolution of 2^j. **Some of the useful properties of Wavelet Transforms are:**

- A high resolution scale decomposition is possible using high-pass and low-pass coefficients
- Transformed data are more compact than original. Noise is distributed throughout the signal while signal in a few coefficients, hence effective data filtering is possible
- Noise models are very well modeled in wavelet space, de-noising can be effectively carried in wavelet space

5 Multi-Resolution CLEAN (MRC)

5.1 Towards Multi-resolution CLEAN

CLEAN is optimized for sources that are not very extended compared to the size of the synthesized beam. Problems will arise if the objects are not well resolved. Högbom introduced the CLEAN method under the assumption that the sky is essentially empty with only a few small sources scattered around. This can looked at in MRA terms as follows:

There is a resolution at which sky appears essentially empty. This implies a hierarchy in the source structure for a collection of varying objects at a similar distance or a collection of similar objects at a varying distance. If a map contains the details and we smoothen it, then the details will be lost but the other structures show up. The latter covers only a part of the sky, so we can CLEAN it up with a fewer parameters. Thus we can break up a given image into different resolutions and CLEAN the Image at each of these resolutions to extract the presence of sources at each resolution. This is the basic idea behind Multi-Resolution CLEAN (MRC) (See Wakker & Schwarz 1988).

MRC doesn't modify the basic CLEAN algorithm but instead it separates the process into several steps. At each step a relatively simple CLEAN is performed with parameters optimized to give a good source intensity estimate. The intermediate step involves decomposing the image into its smoothened and detailed versions by passing through low-pass and high-pass filters. The smoothened data is taken through further steps of decomposition. We can show that for a noise free data if the intermediate steps of MRC are properly combined, then the delta functions found by MRC converges to the delta functions obtained through CLEAN.

One can see that the decomposition used in MRC is similar to Mallat's Algorithm in the wavelet based MRA (see Mallat 1999). Thus wavelets act as effective agents in carrying out MRC. Despite this there is another important reason for the use of Wavelets in deconvolution. After wavelet decomposition the transformed data are more compact than the original signal and thus most of the useful data is distributed in a few coefficients and noise is spread through out all the bands. We can effectively use this fact for de-noising and data filtering. Noise can be very well modeled in the Wavelet Domain and this further adds to its usefulness in de-noising. We should note that low signal to noise had drastic effects in normal CLEAN. This problem is circumvented in the MRC.

The general concept involved in the multi resolution CLEAN is shown in fig.1. We see how the distortions caused by the PSF, transform the true map to a dirty map. The true map can almost completely be retrieved by the MRC.

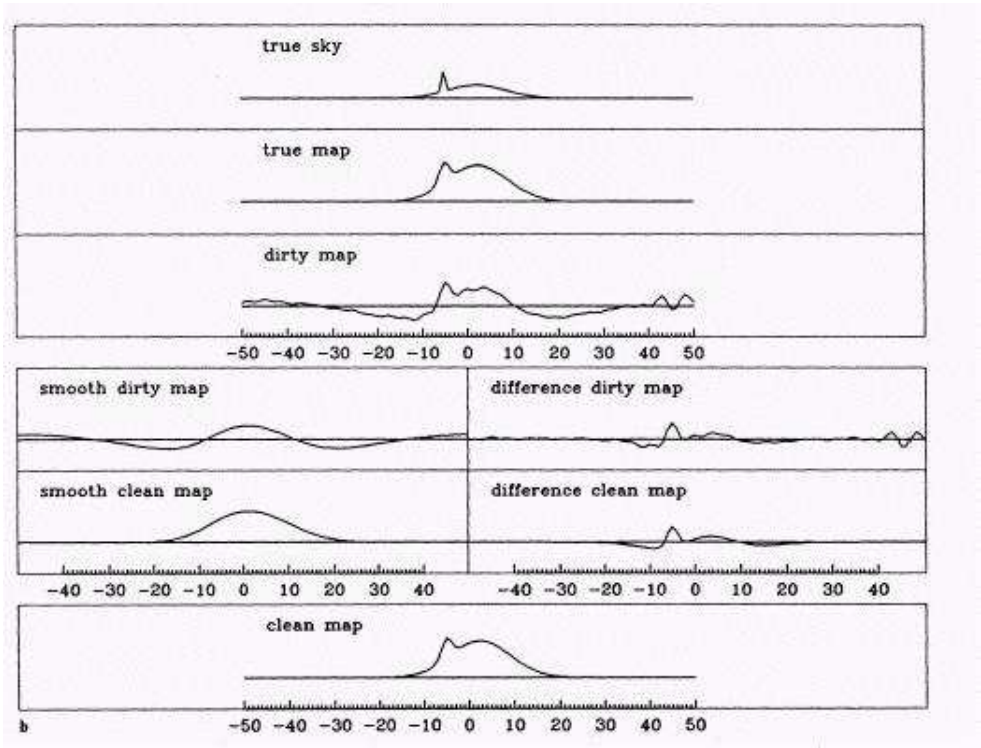

Figure 1: Illustrating MRC in a 1-Dimensional signal with two sources very close to each other (dirty map). The bright source is cleaned through the smoothed map and the difference map cleans the closely located weak source. (Courtesy Wakker & Schwarz 1988)

5.2 The Multi-Resolution CLEAN

Let Φ(x,y) be the scaling function and I(x,y) be the dirty map. Let the caps on the variables denote the Fourier transform of the corresponding variable. The Fourier domain is represented in the u-v plane. The filter coefficients can be found using the dilation equations as follows.

$$\hat{h}(u,v) = \frac{\hat{\Phi}(2u,2v)}{\hat{\Phi}(u,v)} \tag{6}$$

The high pass coefficients can be obtained as,

$$\hat{g}(u,v) = 1 - \hat{h}(u,v) \tag{7}$$

The corresponding wavelet is given by,

$$\hat{\Psi}(u,v) = \hat{g}(\frac{u}{2},\frac{v}{2})\hat{\Phi}(\frac{u}{2},\frac{v}{2}) \tag{8}$$

Our choice of the scaling function was cubic B-Spline* in the frequency domain. First we chose an isotropic (Circular symmetry) B-Spline. [1]

$$\hat{\Phi}(u,v) = B_3(r_{u,v}) \tag{9}$$

where $r_{u,v} = \sqrt{u^2+v^2}$ and $B_3(x) = rect(x) * rect(x) * rect(x)$, where * represents convolution. The block diagram in fig.2 illustrates the steps involved in MRC

We followed the following algorithmic steps for the Multi-Resolution CLEAN.
First we compute the *Wavelet decomposition* as below.

1. Compute the FFT of the image I(x,y) and call it T_0(u ,v)
2. Let n_p be the number of stages of decomposition.
3. Set i=0 and start iteration.
4. Define W_{i+1} as the product of T_i and $\hat{g}(2^i u, 2^i v)$(Note W_0 is zero Matrix) W_{i+1}(u,v)= T_i(u,v) $\hat{g}(2^i u, 2^i v)$.The inverse Fourier transform of W_{i+1} gives the wavelet coefficients at resolution I (scale 2^i)
5. Multiplying T_i by $\hat{h}(2^i u, 2^i v)$ we get T_{i+1} i.e. T_{i+1}(u,v)= T_i(u,v)$\hat{h}(2^i u, 2^i v)$. The inverse Fourier transform of T_{i+1} gives the image at resolution i+1.
6. Increment i
7. If $i \leq n_p$ then go to step 4.
8. The set $w_1, w_2, ..., w_{np}, C_{np}$ gives the wavelet coefficients of the Image. C_{np} is the image at lowest (average) resolution.

We decompose the Image, PSF and Clean beam (i.e. Gaussian beam) by using the above algorithm.

The MRC is performed as shown below;

[1] *B-Spline are auto convolution of the rect-function.

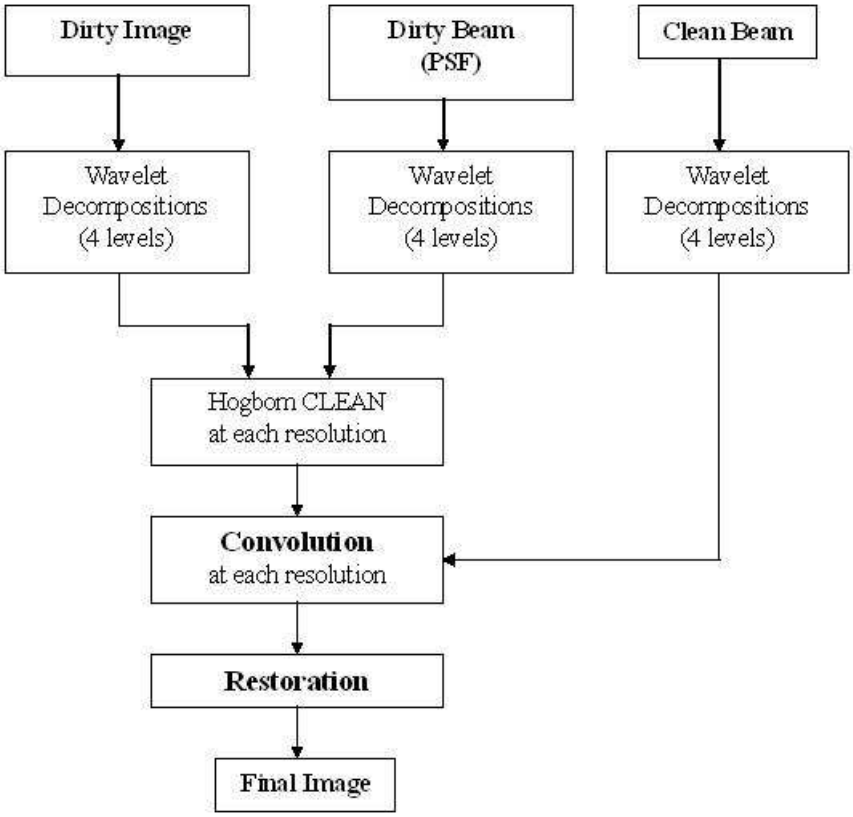

Figure 2: The block diagram of the Multi-resolution CLEAN

if $W_j^{(I)}$(u, v) is the wavelet coefficients of the image at the scale j. Then we get $W_j^{(I)}$(u, v)=$W_j^{(p)}$(u, v)$\hat{O}(u,v)$, where$W_j^{(p)}$(u, v) are the wavelet coefficients of the PSF at scale j. The wavelet coefficients of the image I are the convolution product of the object O by the wavelet coefficients of the PSF.

At each stage j, the wavelet plane $W_j^{(I)}$(u, v) can be decomposed by CLEAN into a set, noted δ_j , of the weighted δ functions.

$$\delta_j = \{A_{j,1}\delta(x - x_{j,1}, y - y_{j,1}), A_{j,2}\delta(x - x_{j,2}, y - y_{j,2}), ...A_{j,n}\delta(x - x_{j,n}, y - y_{j,n})\} \quad (10)$$

Here n_j is the number of δ functions at the scale j and $A_{j,k}$ represents the height of the peak k at the scale j.

By repeating this operation at each scale, we get a set a set W_δ composed of weighted δ-functions found by CLEAN. If **B** is the ideal PSF i.e, the clean beam.Then the estimation of the wavelet coefficients of the object at scale j is given by:

$$W_j^E(x,y) = \delta_j * W_j^B(x,y) + W_j^R(x,y) W_j^E(x,y) = \sum_k A_{j,k} W_j^B(x - x_{j,k}, y - y_{j,k}) + W_j^R(x,y) \quad (11)$$

where W_j^R(x,y)is the residual map.

The reconstruction algorithm obtains the clean map at full resolution.***The Wavelet Reconstruction Algorithm***:

Let the reconstruction filters be $\bar{h}(u,v)$ and $\bar{g}(u,v)$

1. Compute the FFT of the image at low resolution
2. Set j=n_p. Start the iteration
3. Compute the FFT of the wavelet coefficient W_j at the scale j
4. Multiply the Wavelet W_j(u, v) by $\bar{g}(u,v)$
5. Multiply the image at low resolution C_j by
   $barh(u,v)$
6. The Inverse Fourier transform of $W_j\bar{g}(u,v)$+$C_j\bar{h}(u,v)$ gives the image C_{j-1}(x, y)
7. Set j=j-1 go to step 3

The MRC Algorithm in the wavelet space can be summarized as follows;

1. Compute the Wavelet transform of the dirty map, dirty beam and the clean beam
2. CLEAN the dirty map at a particular resolution by the corresponding resolution of the PSF. This decomposes the dirty map into a set of weighted δfunctions i.e.,δ_j.
3. At each scale j, convolve δ_j by the wavelet coefficients of the clean beam and add the residual map W_j^R (this is optional) to the result to obtain the wavelet coefficients of the clean map.
4. Compute the clean map at full resolution by the reconstruction algorithm.

6 The Multi-Resolution CLEAN Experiment

The MRC was performed and confirmed using the follwing procedure. First a test image (see sec 6.1) was generated using the Molonglo Radio Catalogue. Then the PSF of the MRT radio telescope was modelled at -32.5^o, corresponding to the region of the sky taken from the catalogue. Then a clean gaussian beam was generated. The image was CLEANed using the conventional Högbom CLEAN. After that the MRC was applied as described in the previous sections. The results were strikingly different.

6.1 Generating a Test Image

We were developing a new technique for the deconvolution of MRT images. Thus it becomes necessary to test and confirm the technique being developed. For this reason we generated a corrupted (convolved) test image. This was done as follows:

1. We read the radio source information from the Molonglo Radio Catalogue. This was just a testing stage, so we confined ourselves to the Right Ascension range between 18 and 19 hours and a declination between -27.5^o and 52^o.
2. With this information we generated a map of all the sources in this range.
3. This map was divided into five zones. Each of these zones was then convolved with the PSF centered at that zone.

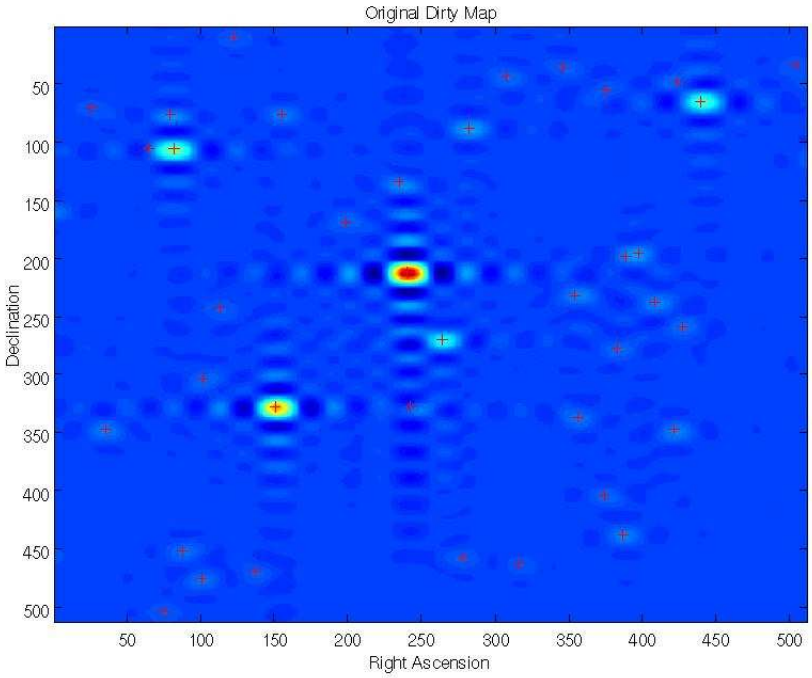

Figure 3: The generated dirty-map

4. The zone pertaining to one particular PSF extended to other zones as well. So it was necessary to merge the different zones after convolution with the proper weightage to each part. This merged image was the dirty map we used for analysis.

The fig.3 shows the generated dirty image.

6.2 Conventional CLEAN

The fig.4 shows the image deconvolved using the Hgbom CLEAN. It was observed that the number of iterations required for the algorithm to converge was considerably large. Now when we lower the loop gain γ the number of iterations increased to a large extent. As we lowered the threshold we could detect more sources. But the lowering of the threshold also resulted in a increase in the number of iterations.

We feel that the large iterations required are a consequence of the presence of a variety of objects with different structure and details. Thus if the sources are extended the CLEAN takes an extremely long time to converge and even then detection is not confirmed. Another important issue concerning CLEAN is the signal to noise ratio. The real images that we deconvolved using the Hgbom CLEAN revealed that the number of iterations required was more. In order to detect all the sources we lowered the threshold. But this also caused certain sidelobes to be detected as radio sources. This certainly is an undesirable effect.

6.3 CLEAN Using Multiresolution

The following is the CLEAN that was performed using Multiresolution. The algorithm as described in fig.2 was perfomed. Figs 5 to 8 show the different resolution of the dirty image in fig 3.Figs 9 and 10 show the 1st and 3rd resolution of the PSF. Figs.11 and

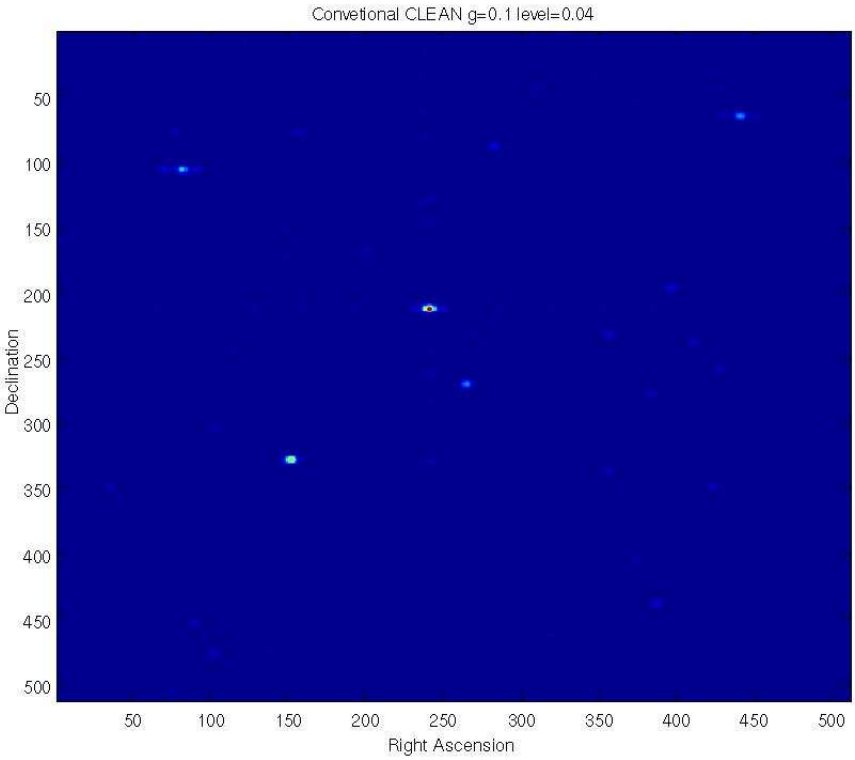

Figure 4: CLEANed using Hgbom CLEAN

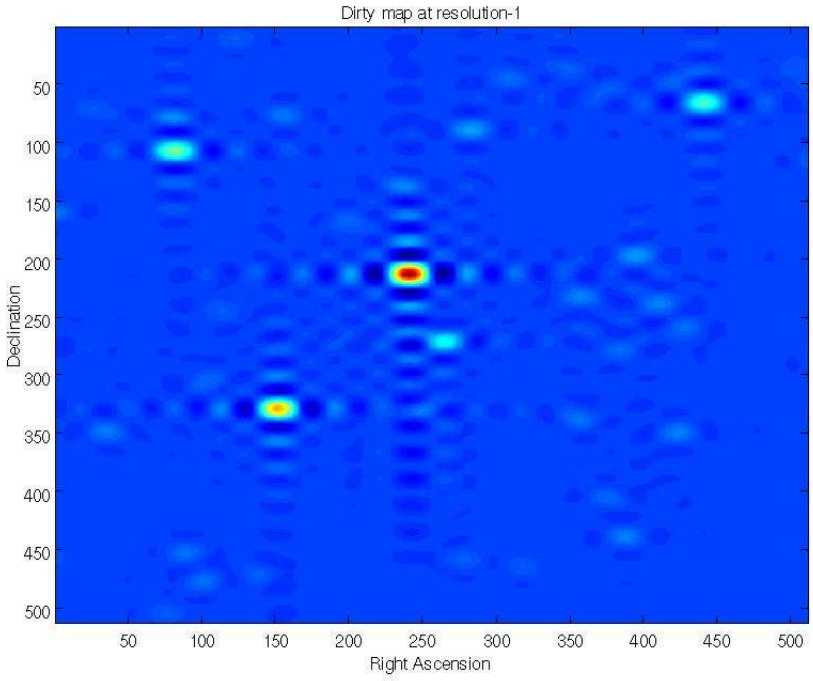

Figure 5: Dirty map at 1st resolution

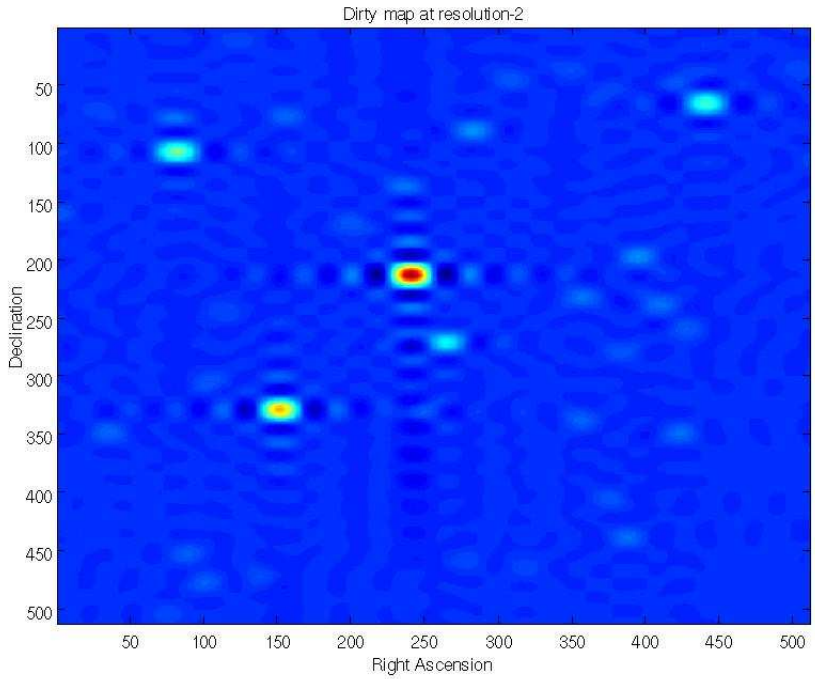

Figure 6: Dirty map at 2nd resolution

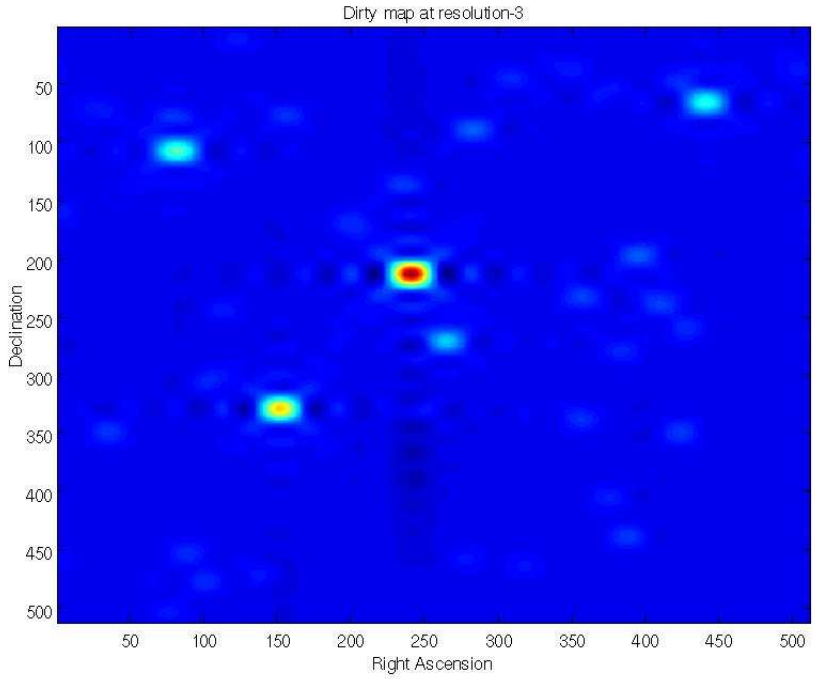

Figure 7: Dirty map at 3rd resolution

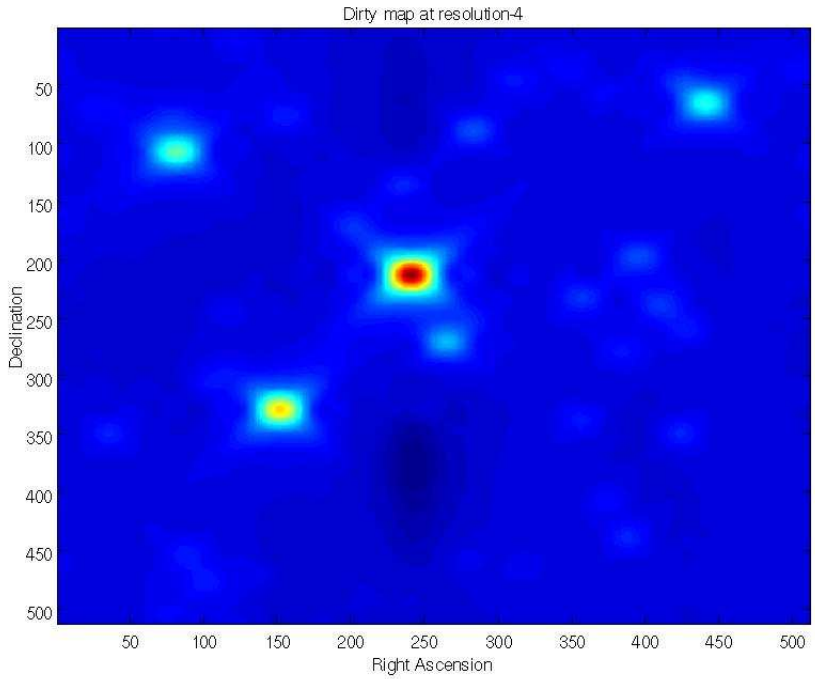

Figure 8: Dirty map at 4th resolution

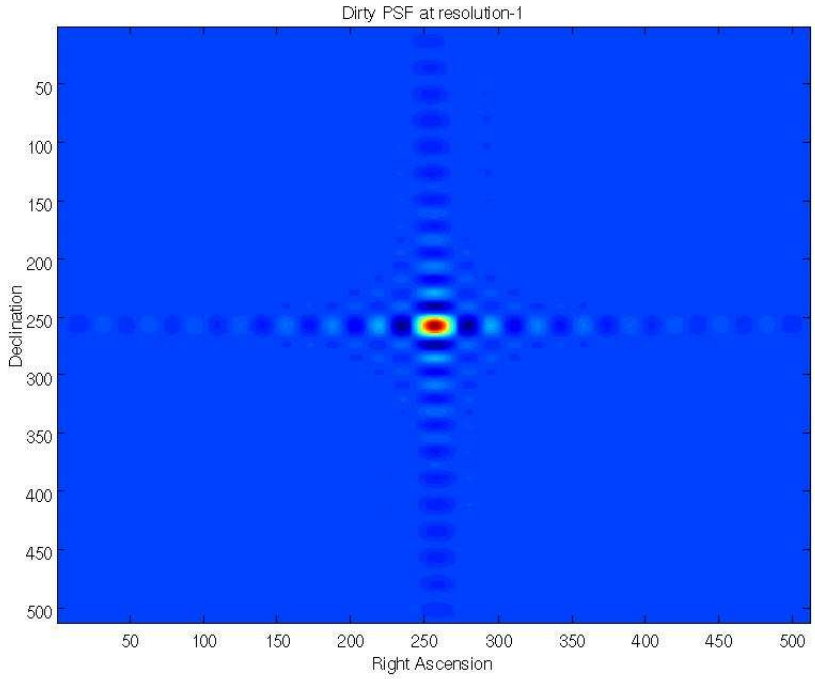

Figure 9: PSF at 1st resolution

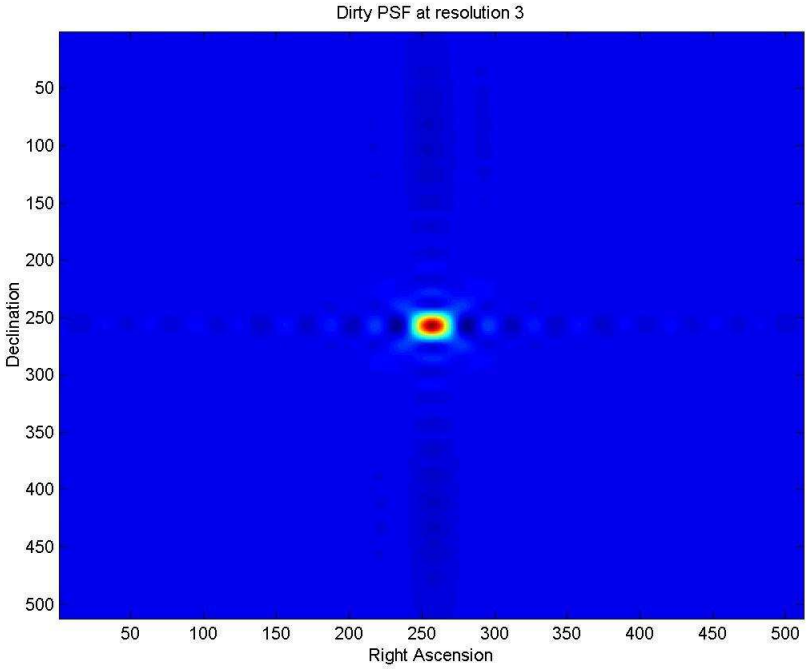

Figure 10: PSF at 3rd resolution

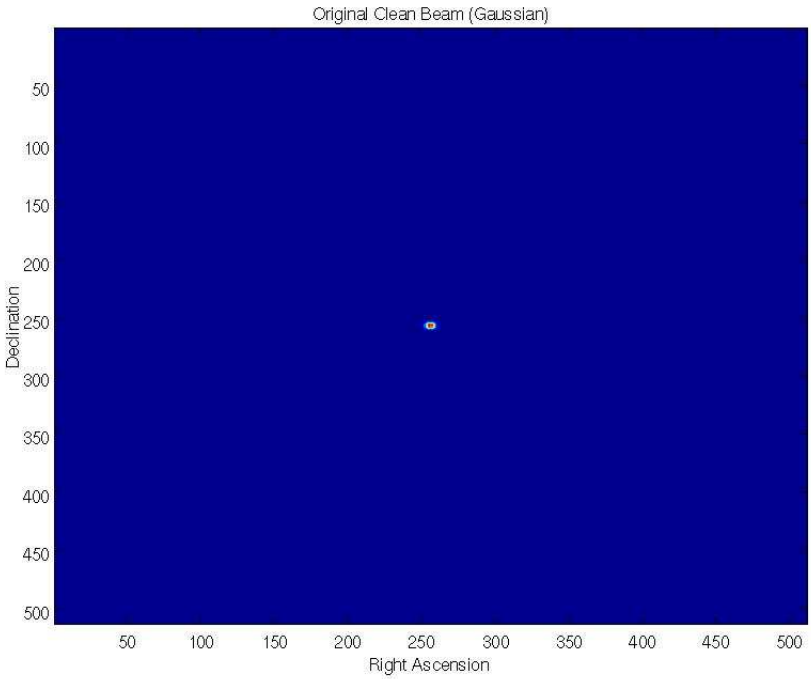

Figure 11: CLEAN Beam 'original'

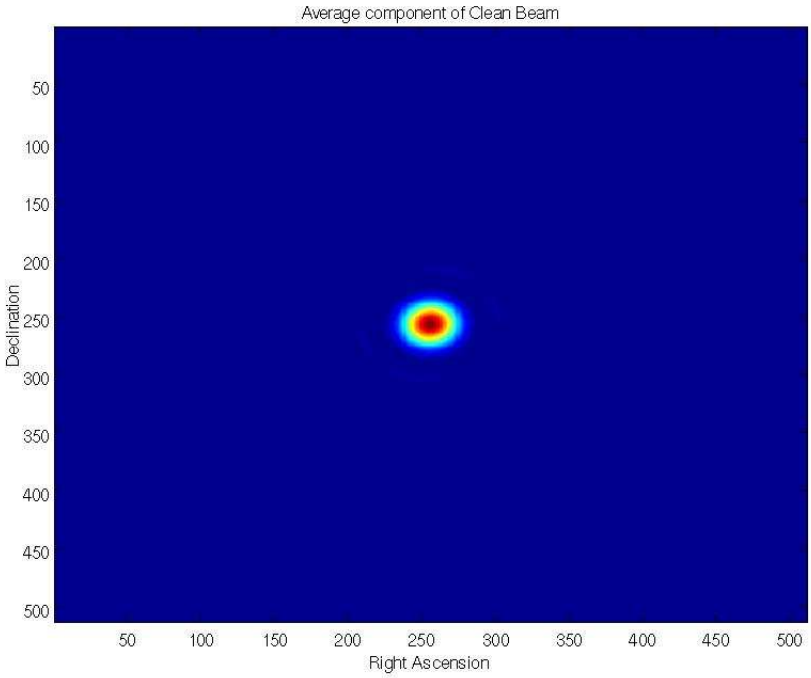

Figure 12: CLEAN Beam 'average resolution'

12 show the CLEAN beam and its average resolution.

Figs 13 to 15 below show the CLEANed Images at different resolutions. We observe that as we go to coarser and coarser approximation the gaussian beam spreads out. This beam would be convolved with the sources detected at different resolutions.

When we run the Multi-Resolution CLEAN (MRC) program at each resolution, we observed that the number of iterations required to detect the sources were considerably less when compared to the conventional CLEAN. To detect a set of sources, the thresholds required by MRC was much less compared to the threshold for normal CLEAN (approximately by a factor of 10). Also we note that, not all sources are detected at each resolution. Different structural details are shown up in different resolutions. Hence deconvolution was carried out effectively at each resolution with lesser iterations.

When we recombined these images at different resolution as described in the algorithm we obtained the final CLEANed map, fig.16,and we see that compared to the conventional clean its performance is much better.

We worked with real image containing significant noise. We found that the MRC was effective in detecting sources even in the presence of this noise. The SNR varied widely in different resolutions. This is an inportant point to note.The lower resolutions had higher signal to noise ratio, and hence we were able to detect the sources at those resolutions without enhancing the sidelobes.Thus in an improved MRC algorithm we could also give weightage to the appropriate resolution. But in order to that we would require some apriori knowledge of the sky.

Thus what we see here is that we have not modified the basic CLEAN algorithm, but have applied it at different resolutions. That fact that the sky looks different and has different features at these resolution makes a difference in the final result.

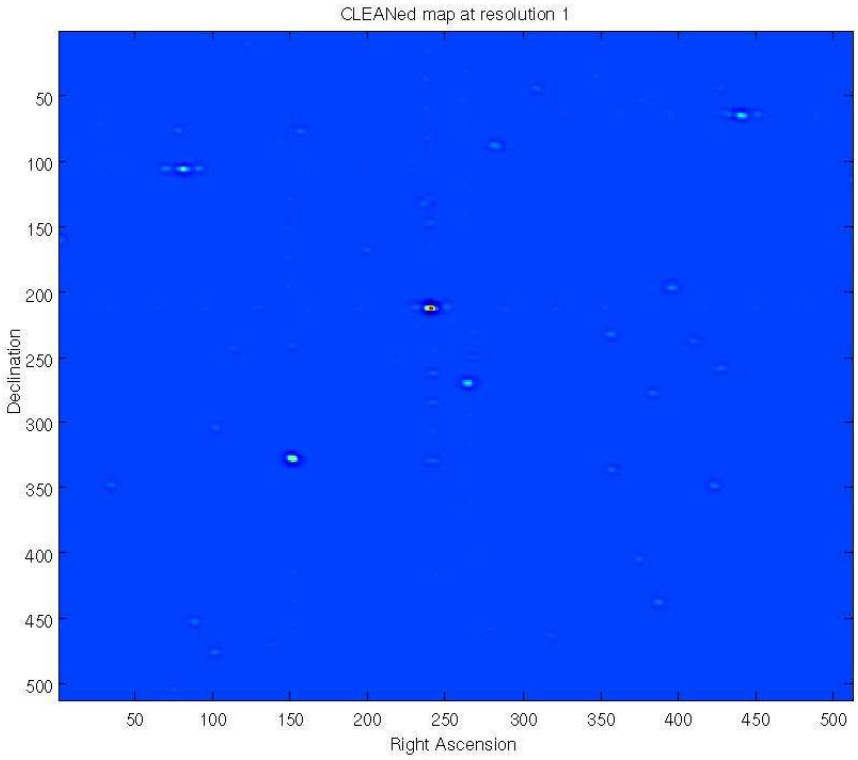

Figure 13: CLEANed map at 1st resolution

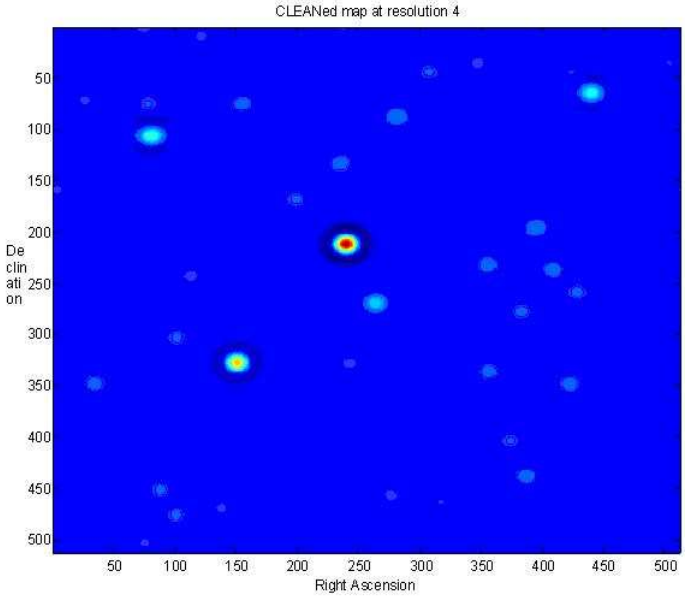

Figure 14: CLEANed map at 4th resolution

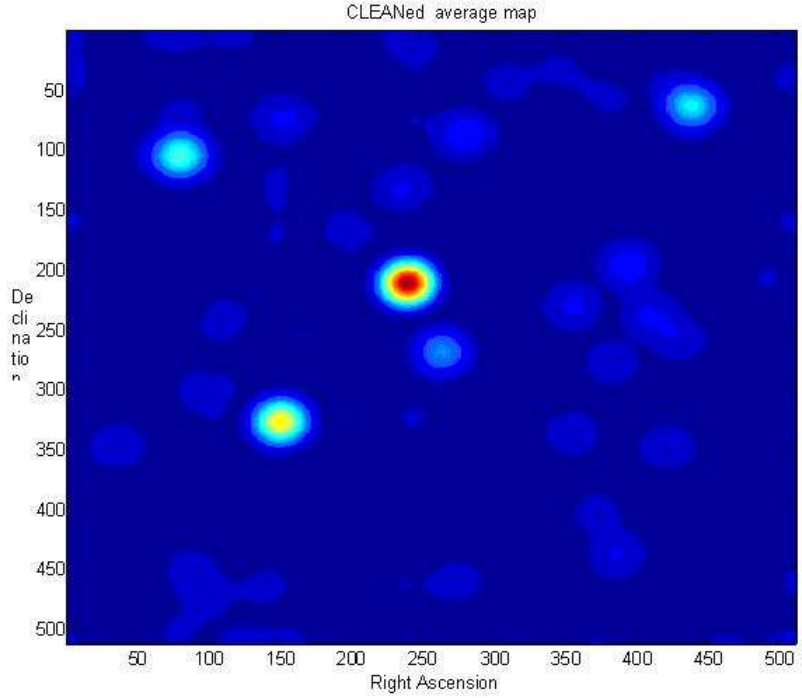

Figure 15: CLEANed map at average resolution

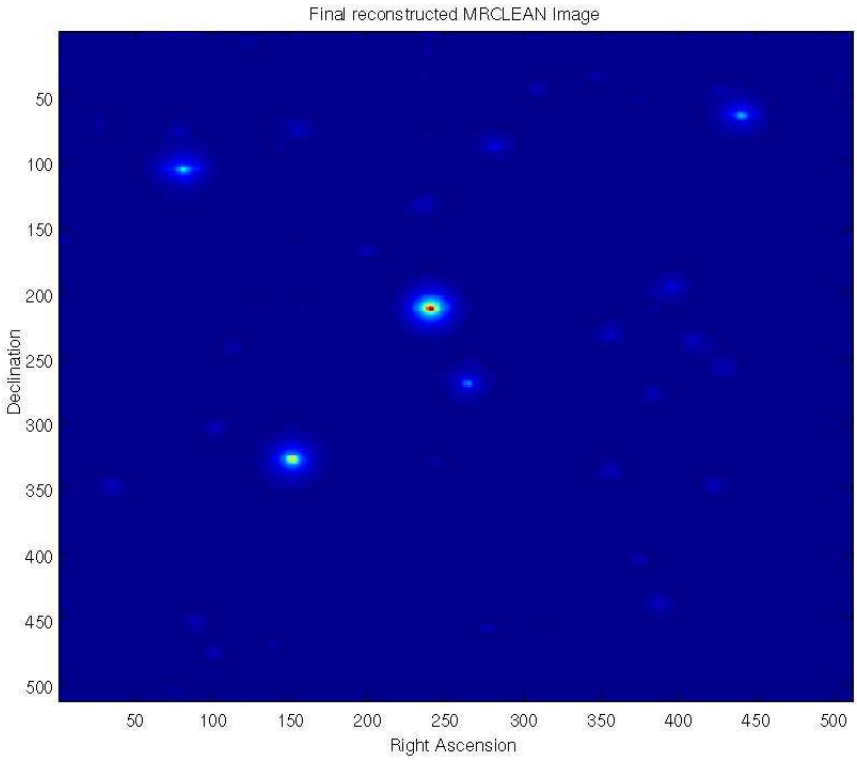

Figure 16: Final CLEANed Map

7 Results and Conclusion

Whenever we are working with wavelets for a particular application, the choice of the scaling function and in turn defining the wavelet becomes a crucial point. The critical parameters which influence our choice are:

1. Shape of the Point Spread Function
2. Support of the Wavelet
3. Number of vanishing moments
4. Regularity

Our scaling function should be able to closely match the PSF.

This facilitates a faster and more effective Deconvolution. In fact many people simply use PSF as the scaling function. Support of the wavelet is important in the sense that a compact support will always aid in digital implementation (FIR Filters). Number of Vanishing moments will determine the regularity or smoothness of the wavelet filter in the time domain.

For our purpose we have used a cubic B-Spline as the scaling function in frequency domain since it closely resembled the PSF 1-D Cross section. We extended it to 2-Dimesion, which resulted in a circularly symmetric scaling function as described by Stark (See Starck et al.,1994).

7.1 Multi-Resolution CLEAN of MRT Images

We extended our implementation of MRC to real images from the MRT data. Fig.17 and fig.18 shows the dirty-map and fig.19 and fig.20 the deconvolved version of it.

7.2 Summary, Conclusion and Scope

Our project, in brief, comprised of understanding the fundamental problems involved in image capturing techniques employed in radio astronomy and implementing a newly proposed technique to counter some of the adverse effects of instrument limitations. The sidelobes of the PSF of an antenna either hid some sources or indicated false sources.

When the sources that are being viewed are extended, the conventional techniques used in radio-astronomy failed to give the correct result. In order to tackle this problem Radio Astronomers suggested a multi resolution approach to eliminating the sidelobes. This meant that the sky would be looked at different resolution and then the images at each of these resolutions would be deconvolved with a fewer number of parameters.

In our project we implemented the multi resolution for a test image (Ref. Sec. 6) and could detect the sources successfully. Specifically, we implemented the algorithm that was given in Starck et al (1994). The scaling function used to generate the filters was a cubic B-Spline. The multi resolution clean on a real map from MRT was also performed and sources detected. This result speaks of the ability of multi resolution CLEAN to clean images with significant noise components.

As a continuation of the work, we would like to apply the multi resolution clean to all the images from MRT. We also feel that scaling functions that depend on the PSF of the antenna array could produce a better result.

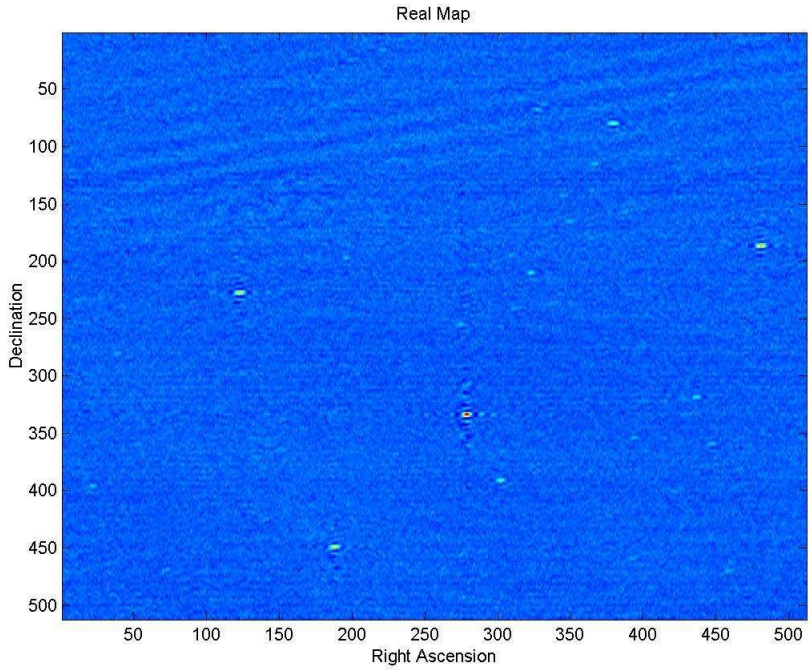

Figure 17: MRT dirty map

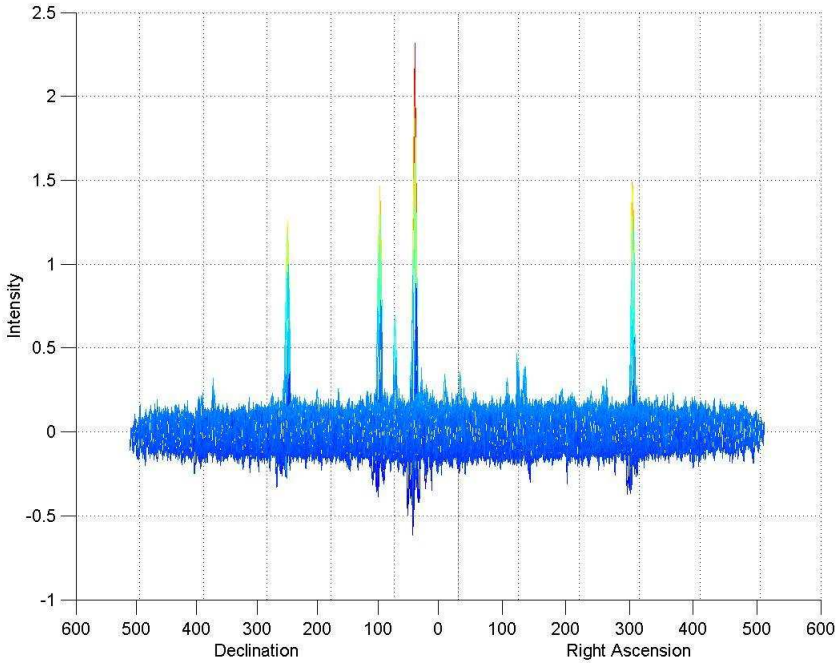

Figure 18: MRT dirty map

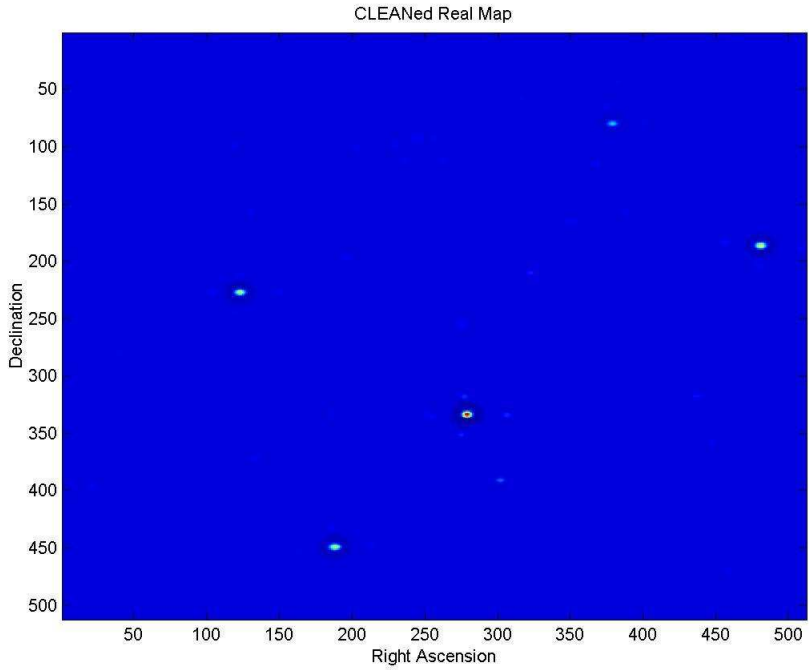

Figure 19: MRT CLEANed Map

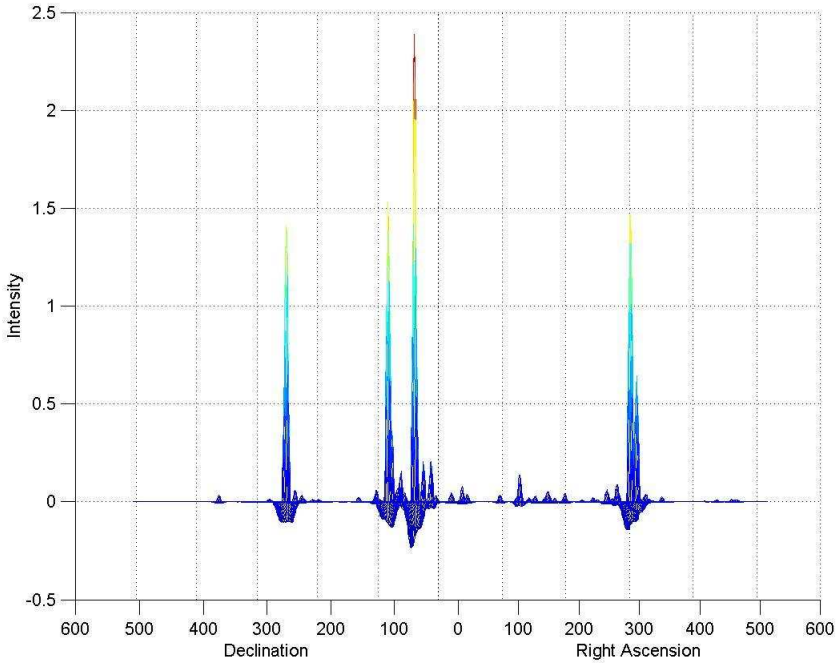

Figure 20: MRT dirty map

We were successful in retrieving all the point sources we started with provided by the Molonglo Radio Catalogue. All the sources were detected with a lower number of iterations compared to the conventional CLEAN.

As a continuation of our efforts, we are trying to extend the algorithm to real images with extended features.

We used the same relative threshold level for both conventional as well as Multi Resolution CLEAN and it can be seen from the final cleaned images of both the techniques that the sidelobes are visible in the conventional CLEAN but not in the Multi-Resolution CLEAN.

We were successful in deconvolving the test image we started with. The entire sources that were used to generate the dirty map were detected. The algorithm was also used on a real image and we obtained satisfactory results. But in order to get a complete measure of the usefulness and the power of multi resolution CLEAN it is necessary to apply the MRC to the entire MRT image. Also, more tests have to done on images with extended sources and noise.

Acknowledgements

This study was made possible by the support of the Raman Research Institute, India and Chalmers University of Technology. I thank the authorities in these institutions for providing me with the facilities.

I especially would like to thank Prof. Udaya Shankar and Prof. Alessandro Romeo for their advise and technical support during the project. The research team from Mauritius was very helpful in getting us aquainted with many of the already existing MATLAB routines.

Finally I should also thank my colleagues, especially Shivakumar Jolad (University of Texas A & M) for his contribution to the project. Also Magnus Franzen of the Master's program in Radio Astronomy helped me with the documentation in latex, I am thus grateful for his support.

References

Bhat N.D.R, Cordes J.M., Chatterjee S., 2003, ApJ, 584, 782.

Burke Bernard F., Smith Francis Graham, 2002, An Introduction to Radio Astronomy, Cambridge: Cam. University Press.

Cornwell, T.J., 1983, A&A, 121, 281

Högbom, J. (1974), ApJ. Suppl. Ser., 15, 417.

Jansson Peter A., 1997, Deconvolution of Images and Spectra. San Diego:Acad. Press.

John D. Kraus, 1986, Radio astronomy, 2nd Edition, Powell Ohio : Cygnus-Quasar Books.

Mallat Stephane, 1999, A Wavelet Tour of Signal Processing. London: Acad. Press.

Narayan, R., Nityananda, R., 1986,Ann. Rev. A&A., 24, 127.

Siddhichai Supakorn., Chambers Jonathan, 2002, "Wavelet-based blind image deconvolution," presented at the colloq' on Sig.Proc, Thailand.

Starck Jean-Luc., Bijaoui Albert, 1994, Sig. Proc, 34, 195

Starck Jean-Luc, Bijaoui Albert., Lopez Bruno.,Christian Perrier., 1994, A&A, 283, 349.

Schwarz, U.J., 1978, A&A, 65, 345.

Wakker B.P, Schwarz U.J, 1988, A&A, 200, 312.

Wakker B.P, Schwarz U.J, 1990, in "Radio Interferometry: Theory, Techniques, and Applications: Proceedings of IAU Colloq' 131, Socorro, NM,(A92-56376 24-89). San Francisco, CA,Astronomical Society of the Pacific, 268.

Dark Matter in the Universe and Alternatives

Farhad Aslani

Göteborg University
SE-41296 Göteborg, Sweden
(gu97faas@dd.chalmers.se)

*

Abstract

Dark matter (DM) and what legitimates the existence of it is discussed. We talk about flat rotation curves of galaxies as well as large scale structures in the Universe, like galaxy clusters and superclusters, gravitational lensing, and fluctuations in the cosmic microwave background (CMB) radiation. The Tully-Fisher relation is explained. Some DM candidates, both baryonic and non-baryonic, are mentioned briefly. Newtonian dynamics and modified Newtonian dynamics (MOND) are discussed. Finally, DM and MOND theories are compared.

1 Introduction

What would you say if I told you that 98% of all things around you are invisible? For many astronomers, this is an everyday issue. They observe phenomena which can be explained by assuming that there is much more invisible, or so far undetectable, matter in the Universe than there is visible matter. This kind of matter is called dark matter, because it does not reflect or illuminate any considerable light (if any at all). There are, however, alternative explanations to the phenomena mentioned above. In the following text we will talk about these phenomena, dark matter theory, and these alternatives.

2 Celestial Mechanics

Here we give a quick reminder of Newtonian mechanics, which is used frequently in astronomy. The special or general theory of relativity is not applied for our purpose since we have low velocities and accelerations as well as weak gravitational fields. However, we will briefly mention general relativity in connection to gravitational lensing.

When you want to describe the motion of celestial bodies and galaxies in the Universe, the gravitational force, which is almost the only significant force on large scales, has the following expression in the case of two bodies:

$$F = \frac{GMm}{r_{12}^2}, \tag{1}$$

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

where M and m are the masses of two bodies, r_{12} the distance between their centers of mass, and G is Newton's gravitational constant. If $M \gg m$, It is convenient to the take body with mass M to be stationary. Then the equation of motion for the body with mass m is

$$-\frac{GMm}{r^2}\hat{\mathbf{r}} = m\ddot{\mathbf{r}}. \tag{2}$$

For circular motion of m around M, solving equation (2) gives:

$$\frac{GMm}{r^2} = mr\omega^2 = m\frac{v^2}{r}, \tag{3}$$

with angular velocity ω, or linear tangential velocity v. r is the radius of the circle. Equation (3) is the well known expression for equilibrium between centripetal and gravitational forces. Taken together, these equations yield

$$v = \sqrt{\frac{GM}{r}}. \tag{4}$$

For elliptical motion of m around M, we give the expression without going into details:

$$\sqrt{\frac{GM}{r_{\min}}} < v < \sqrt{\frac{2GM}{r_{\min}}} \tag{5}$$

If the velocity is greater than $\sqrt{2GM/r_{\min}}$, the two bodies will not be connected to each other gravitationally.

3 Observations

Here we review the observations of phenomena which have led to the invention (or discovery) of the idea of dark matter. In some literature these observations are referred to as evidences, but I choose to not use this word.

3.1 Rotation Curves of Galaxies

One of the observations which has contributed to the idea of dark matter is the velocity dispersion (that means how velocities change with radii) of stars rotating around the Galactic center. We can approximate the motion of stars around the Galactic center, with circular or elliptical motion of one body with low mass around another with much higher mass. Then we can use eq.(4) or eq.(5) to describe the motion. Here $M = M(r)$ is radius dependent, and increases with r.

Due to the rapid increase of $M(r)$ in the more central regions of the Galaxy, the velocity for circular motion will increase when you go away from the galactic center, which is confirmed by observations (see figure (1)). When we go more and more outwards, the Galaxy is rather sparse and we don't have any significant growth of $M = M(r)$. Then from equation (4) it is expected that the velocity is related to r as $r^{-1/2}$. But observations shows something completely different (see figure 1). Figure (2) shows the expected velocity dispersion. Obviously there is a huge discrepancy between the observed and calculated velocity dispersions for this galaxy. In the outermost regions, v_{obs} is almost constant, whereas v_{cal} decreases as $1/\sqrt{r}$. It is said that we have a flat rotation curve.

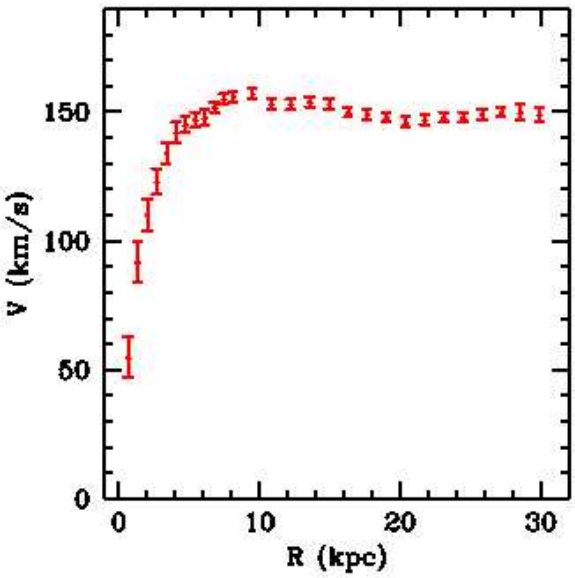

Figure 1: Observed velocity dispersion of the galaxy NGC3198. Increasing velocities for $R < 10$kpc, almost constant velocities for $R > 10$kpc. From Begeman, 2001, University of Berkeley.

Similar results are obtained for many other galaxies, and if we assume equation (4) to be valid, then $M(r)$ must increase as $M(r) \sim Cr$, where C is a constant. This is in contradiction with what we "see"; a sparse space around the galaxies.

This contradiction can be explained, by assuming that there is a lot of matter in galactic halos which we cannot observing from electromagnetic radiation. Or by introducing the idea of existence of dark matter in galactic halos can we solve the problem. If we find this dark matter some way or other, then the idea is correct, otherwise we shall keep looking or find another explanation. But there are other observations which support the idea of dark matter.

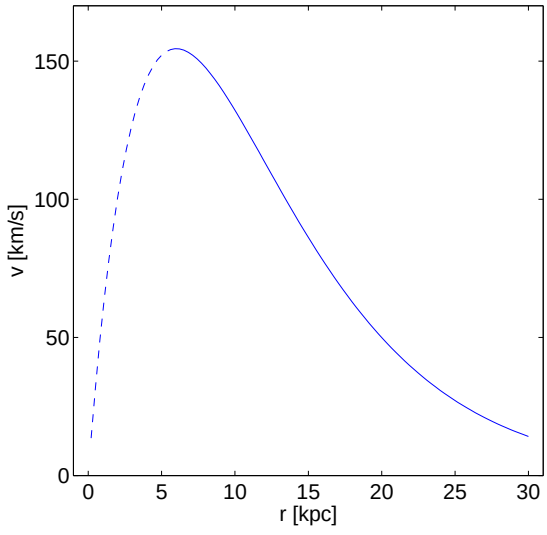

Figure 2: Expected velocity dispersion in a disk galaxy (approximation)

3.2 The Tully-Fisher (TF) Relation

The Tully-Fisher relation is an experimental law which connects luminosity or absolute magnitude of a galaxy to its maximal circular rotation velocity:

$$L \propto v_c^4. \tag{6}$$

Since all galaxies of the same type have roughly the same proportion of certain types of stars, we can assume that the mass of a galaxy is proportional to the absolute magnitude or luminosity; $M \propto L$. Since the surface area is proportional to the square of the radius, $S \propto R^2$, and the luminosity, $S \propto L$, we can write:

$$M \propto L \propto S \propto R^2 \tag{7}$$

If we compare equations (3) and (7), we have:

$$v_c^2 \propto \frac{M}{R} \propto \frac{L}{R} \Rightarrow Rv_c^2 \propto L \Rightarrow R^2 v_c^4 \propto L^2\,, \tag{8}$$

and using eq.(7) once again we get:

$$Lv_c^4 \propto L^2 \Rightarrow L \propto v_c^4, \tag{9}$$

which is the Tully-Fisher relation from equation (6). Measurements of circular velocity goes as follows. If you look at galactic disc from the side (see figure(3)), the light from

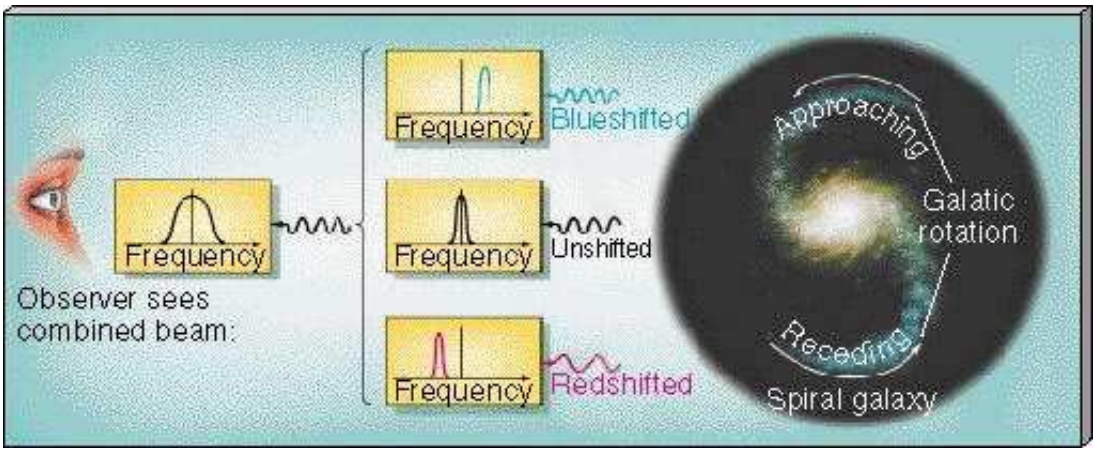

Figure 3: Red and blue shift from a rotating galaxy and Broadening of spectral lines. From Jim Brau, 2001, University of Oregon.

the part coming towards you will be blueshifted, while the part going from you will get redshifted. This will result in broadening of spectral lines. By the studying spectral lines, we can measure v_c for different galaxies. When we know the distance and apparent magnitude, we can calculate luminosity, and see if TF holds. If you look at a tilted galaxy you need to do corrections and calculate v_c. The fact that experiments support the TF relation, shows that there is proportionality between M and L. If we take TF as a law, we can use it to measure the distances to the galaxies.

Figure 4: A galaxy cluster. From the European Southern Observatory (ESO), 1998

3.3 Galactic Clusters and Super-Clusters

Galaxies are not randomly distributed in the Universe, they actually build structures like clusters and super-clusters. A galaxy cluster consist of galaxies which are gravitationally bound, and rotate around their centers of mass. Sometimes several clusters are gravitationally bound together and build super-clusters. With the help of the Tully-Fisher relation we can measure the distances from us to the galaxies. We can also easily measure the apparent distances between the galaxies. then we can calculate the true distances between the galaxies and see if they are close together, i.e. belonging to same cluster. In figure (5) the graph shows the relation between apparent magnitude and velocity of galaxies belonging to two different clusters. From equation (6) we get:

$$L \propto v^4 \Rightarrow L = C_1 v^4 \Rightarrow \log L = C_2 + 4 \log v \tag{10}$$

with C_1 and C_2 constants. We know that the apparent magnitude m is a logarithmic function of L/R^2. Thus in case of equal distances it is a logarithmic function of just luminosity. From this fact and equation (10), we conclude that m is a linear function of $\log v$. Figure (5) shows such a connection.

With the help of Doppler shift we can measure the velocities of galaxies in a cluster relative to us and by that relative to each other. By equations (2) and (6) we can calculate the masses and see if equations (4) and (5) hold. Measurements have shown that the visible mass is too low in clusters to keep their constituent galaxies together. It is also too low to make clusters gravitationally bound to build super-clusters.

3.4 Gravitational Lensing

According to Einstein's theory of general relativity, when light propagates close to a massive body, the path is bent by this body (see figure (6)). If the bending body has

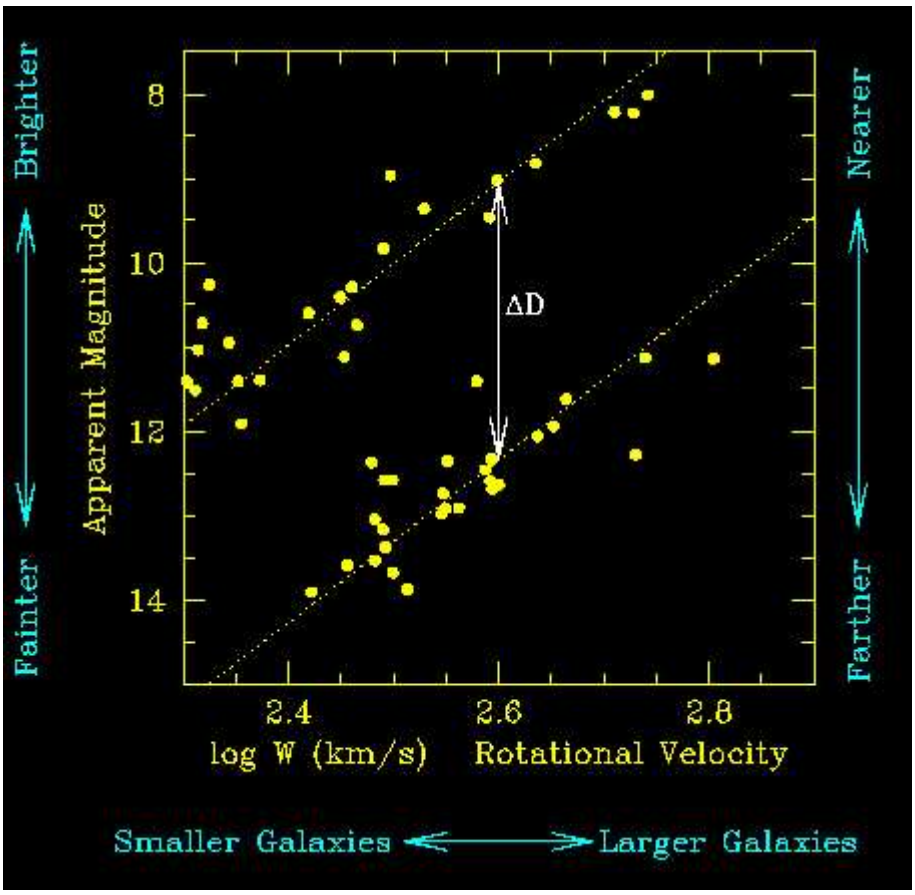

Figure 5: Linear relation between m and $\log v$. From National Optical Astronomy Observatory (NOAO).

higher gravitational acceleration it will bend the light even more. The explanation of this bending phenomenon is rather complicated, and needs 4-dimensional non-Euclidian geometry, but you can think about it as if light is attracted gravitationally towards and by the body. If you put a massive object near the path of light between a light source and an observer, the image of the light source from the observers view will be somewhat distorted. This phenomenon is called gravitational lensing. There are mainly two kinds of gravitational lensing, micro and macro lensing.

In the first case you usually have one single object (it can be a star in some stage of it's evolution) which bends the light of a single light source, usually a bright star. It will lead to apparent displacement of the light source or magnification of it's brightness.

In the second case a whole galaxy or even a galaxy cluster contributes to the bending of light, which comes from other galaxies. It makes these look like arches. In figure (7) you see an example of this phenomenon.

Micro lensing has been observed several times in our own galaxy and is an indication of the existence of baryonic dark matter. Macro lensing phenomena of the kind we see in figure (7), need more mass than the estimated mass of the visible matter.

3.5 Cosmic Microwave Background Radiation (CMB)

Cosmic microwave background radiation comes from the early stages of evolution of the Universe, and is the oldest thing which we can observe. According to the big bang theory, before stars and galaxies were built, the Universe was an almost homogeneous dense

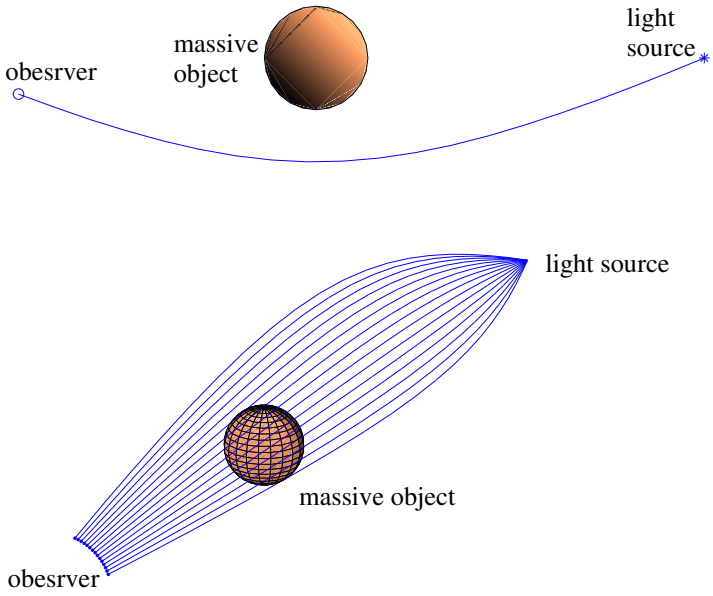

Figure 6: A gravitating body bends the path of light.

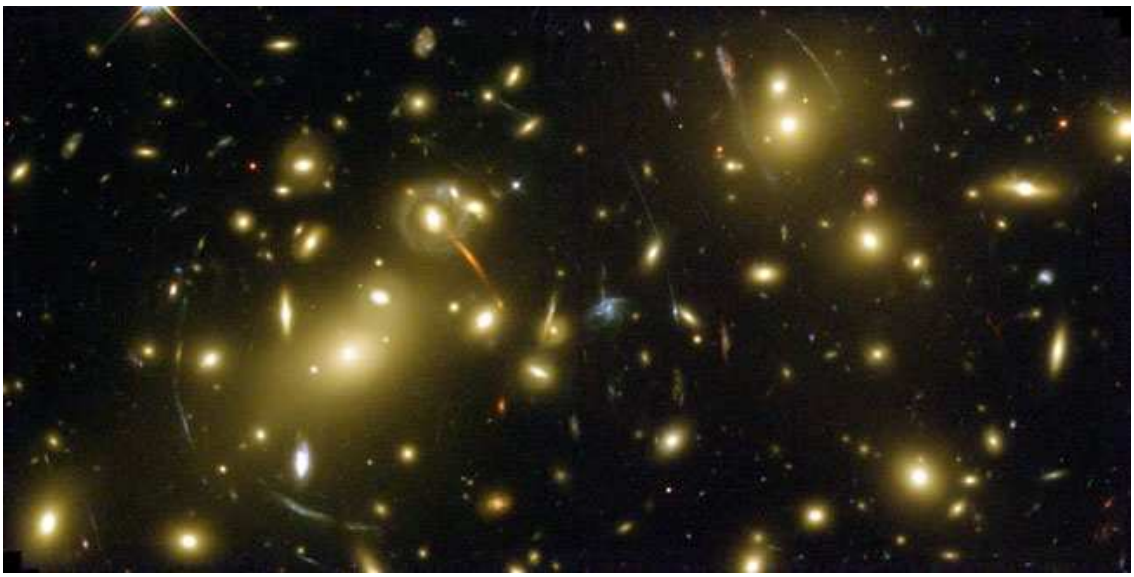

Figure 7: Arches caused by gravitational lensing, from galaxy cluster Abell 2218. From Hubble Space Telescope, NASA.

gas of electromagnetic radiation and baryons, called the photon-baryon fluid. There were however small fluctuations in that gas: it was somewhat denser in some places and thinner in other places. By gravitation, the denser areas attracted, matter from the less dense areas. In other words gravity amplified the already existing fluctuations. Denser areas became denser and thinner areas became thinner. This kept going on, until the first structures of the Universe were constructed. The CMB comes from the time directly before the structures began to build. It tell us how fluctuations in the photon-baryon fluid were distributed, and by studying the CMB we can learn how the first stars, galaxies and clusters were born. Studies have shown that gravity from baryonic matter could not have been sufficient to make the structures in the universe that we see today. You need more than baryonic matter.

4 Dark Matter Candidates

Astronomers denote the density of the Universe by Ω. The unit is decided in such a way that you have the following situations:

$$\begin{aligned} \Omega > 1 &\Rightarrow \text{positive curvature} \\ \Omega = 1 &\Rightarrow \text{no curvature} \\ \Omega < 1 &\Rightarrow \text{negative curvature} \end{aligned} \tag{11}$$

The most recent observation have shown that $\Omega \approx 1$. According to the ΛCDM theory (Λ Cold Dark Matter, where Λ refers to the dark energy), the major part of the energy density in the universe, nearly 70%, is build up of an exotic form of energy, called vacuum energy or dark energy. The remaining 30% is matter, luminous and dark. About 4% of total energy, or 10% of matter, is baryonic mass, in other words objects consist mainly of protons and neutrons. The rest of the matter, approximately 80% of it, is some kind of exotic matter which we don't know so much about. Only about 10% of the baryonic dark matter is visible. We can summarize energy and mass distribution in the Universe as below:

$$\begin{aligned} \Omega_{\mathrm{DE}} &\approx 0.7 \\ \Omega_{\mathrm{m}} &\approx 0.3 \\ \Omega_{\mathrm{nb}} &\approx 0.26 \\ \Omega_{\mathrm{b}} &\approx 0.04 \\ \Omega_{\mathrm{v}} &\approx 0.005 \end{aligned} \tag{12}$$

In this paper we only discuss dark matter.

4.1 Baryonic Dark Matter (BDM)

As you see from equation (12), there is not so much BDM in the Universe, so we talk very briefly about it. Old stars that have burned their fuel and are now cold dark objects can be baryonic dark matter, like black holes, neutron stars or chilled white dwarfs. "Stars" that never have been big enough to be stars, and primordial black holes (black holes who were built in early stages of the Universe long before stars) are other kinds of BDM. All these are called MAssive Compact Halo Objects, MACHOs. Some of them have actually been detected by gravitational lensing.

Another kind of Baryonic Dark Matter is big cold gas clouds, which are too cold to be detected easily. But as equation (12) shows, BDM is not sufficient to account for the considerable amount of the missing mass.

4.2 Non-Baryonic Dark Matter (NBDM)

To understand NBDM, we need comprehensive knowledge about particle physics, which is beyond the scope of this text. Therefore we will give only a very brief introduction.

I use here a classification which Paolo Gondolo has presented, and I find it very instructive:

Type Ia: we know that they exist, for example neutrino.

Type Ib: their existence have been suggested to solve particle physics problems, without having anything to do with dark matter, for example neutralinos and axions. The neutrino was once such a particle: Austrian physicist Wolfgang Pauli had suggested that it existed to solve the energy conservation problem in β-decay.

Type II: is all other: WIMPZILLAs, solitons, super symmetric particles, matter from parallel Universe. Many of them are actually invented only to solve the dark matter problem.

The abbreviation WIMPs, Weakly Interacting Massive Particles, is used frequently for many of NBDM candidates.

There are still questions about how much mass the neutrino has. If it is heavy enough, it can be responsible for a part of dark matter. Another problem is that it is a Hot Dark Matter (HDM) particle, which means that it has high velocity. As we said in section 3.5 about CMB, dark matter has played a significant role in building the first structures of the universe, and studies of the CMB have shown that the major part of dark matter must be low velocity particles or Cold Dark Matter (CDM).

Other candidates have not yet been detected, and some of them may never be detected even if they exist. There are ongoing experiments, both with accelerators on earth and all kinds of radiations from space, to detect them. One problem so far is that accelerators on earth can not yet provide the high energies necessary for creation of particles.

5 Alternatives to Dark Matter Theory

We can summarize the idea of dark matter in a few sentences: "There is more gravitation in the Universe than equation (1) can account for. If M and m are masses of visible matter, then there must be more matter in the Universe than the visible matter". For example for flat rotation curves of galaxies, we have too much v in equations (4), and this tells us that M must be bigger than what we see. (As you know G and r can not be wrong.). But can we be sure that these equations describe correctly the motions of celestial bodies? We have tested Newton's mechanics only in scales as large as the solar system, and we don't know if it is applicable on galactic scales. We know actually that classical mechanics fails to be true when you have things which are as small as or smaller than atoms, or when you have velocities comparable with the speed of light. No experiments have ever shown that the classical mechanics is valid in galaxy scales. In contrary many experiments tell us the opposite.

There are actually people who have suggested other mechanical descriptions to explain our problem with the missing mass. Most of these new descriptions are rather easy to falsify. For example one of them says that eq.(1) is not valid if we have big distances. This can not be true, because flat rotation curves phenomenon occur on different distances in different galaxies.

But one of them have appeared to be a strong challenger to the dark matter solution. It is Modified Newtonian Dynamics (or MOND), proposed by Mordehai Milgrom. Here we will talk more about that.

5.1 Modified Newtonian Dynamics (MOND)

You can get MOND's equation in two ways; either by modifying the gravitational law (equation (1)) or by modifying Newton's second law $F = ma$, which is equivalent to modification of inertia. In the first case we modify the Poisson's equation which for Newton's gravitation law is:

$$\nabla \cdot (\nabla \phi) = \nabla^2 \phi = 4\pi G\rho\,, \tag{13}$$

with ϕ the gravitational potential, ρ density of matter and G Newton's gravitational constant. For bodies with spherical symmetry, it yields eq.(1). The modified version of Poisson's equation is:

$$\nabla\cdot\Big[\mu\Big(\frac{|\nabla\phi|}{a_0}\Big)\nabla\phi\Big] = 4\pi G\rho\,, \tag{14}$$

with characteristic acceleration a_0 and scalar function $\mu(x)$ with the following character:

$$\mu(x) = \begin{cases} 1, & x \gg 1 \\ x, & x \ll 1 \end{cases} \tag{15}$$

Some of the candidates for $\mu(x)$ are:

$$\mu(x) = \frac{x}{1+x} \quad ,$$
$$\mu(x) = \frac{x}{\sqrt{1+x^2}} \quad ,$$
$$\mu(x) = 1 - e^{-x}$$

In figure (8) you see what $\mu(x)$ approximately looks like. If we integrate eq.(14) twice, in

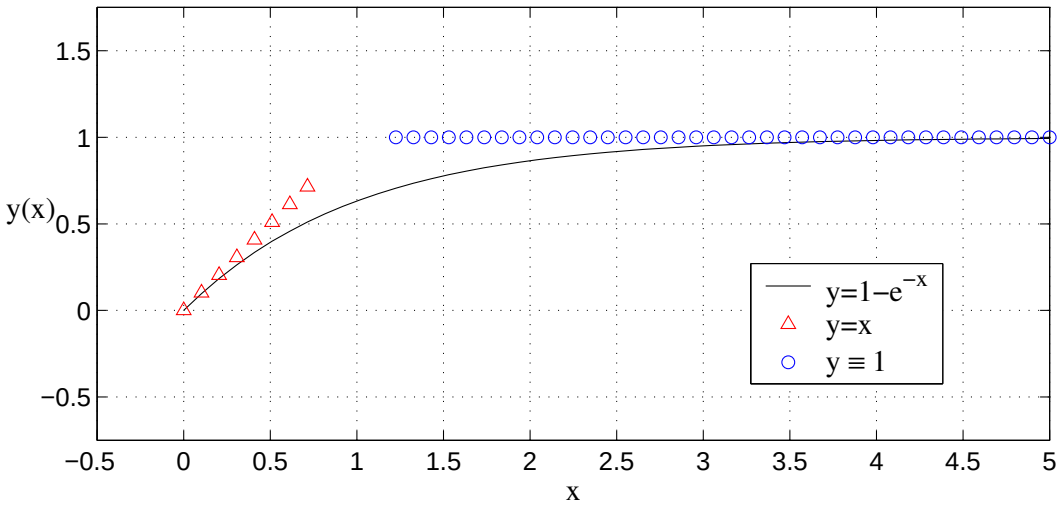

Figure 8: Approximate behaviour of $\mu(x)$.

the case of spherical symmetric mass distribution, we get:

$$\begin{aligned} \int_V \nabla\cdot\Big[\mu\Big(\frac{|\nabla\phi|}{a_0}\Big)\nabla\phi\Big]dV = \int_V 4\pi G\rho dV \Rightarrow \\ \int_S \Big[\mu\Big(\frac{|\nabla\phi|}{a_0}\Big)\nabla\phi\Big]\cdot d\mathbf{s} = 4\pi GM \Rightarrow \\ 4\pi r^2\mu(\frac{a}{a_0})a = 4\pi GM \Rightarrow \\ \mu(\frac{a}{a_0})a = \frac{GM}{r^2} = g_N \end{aligned} \tag{16}$$

Here we have used a for $\nabla\phi$, and g_N refers to Newtonian gravity acceleration.

In the case of modifying the inertia, Newton's second law will be replaced by

$$F = m\mu(\frac{a}{a_0})a\,, \tag{17}$$

with a_0 and μ as before. The characteristic acceleration a_0 should be decided experimentally. The currently accepted value is about 1.2×10^{-12} m/s^2.

It's easy to see from equation (16) that:

$$\begin{cases} a \gg a_0 & a = g_N \Rightarrow \quad g_N \gg a_0 \\ a \ll a_0 & a = \sqrt{a_0 g_N} \end{cases} \tag{18}$$

Also if g_N is sufficiently large, (larger than a_0) we have usual Newtonian acceleration which we observe in everyday life and in the solar system. but if g_N is much less than a_0 we get:

$$a = \sqrt{a_0 g_N} = \sqrt{a_0 \frac{GM}{r^2}} \propto \frac{1}{r} \tag{19}$$

Equating this with the centripetal acceleration from equation (3) yields :

$$\frac{v^2}{r} = a \propto \frac{1}{r} \Rightarrow v = \text{constant}\,. \tag{20}$$

You see that MOND beautifully explains the flat rotation curves of galaxies without the need of any dark matter.

6 Comparison

Which is right? Dark matter theory or MOND? If you ask me, I will answer maybe neither of them! They both have their strong and weak sides. Here we will try to do a comparison between them.

6.1 Problems with Dark Matter

- Cuspy galactic cores: To explain flat rotation curves of galaxies, the dark matter distribution in it must be rather smooth. But simulations have shown that non-baryonic dark matter will be highly concentrated in the center of galaxies, The reason is that dark matter doesn't feel any pressure either from itself or from baryonic matter. Some astronomers have begun to wonder if dark matter can interact with itself, (or as some people say, can dark matter see itself.)

- Dark galaxies or sub halos: Simulations have shown also that there must be many small sub halos in a halo, and thereby many small dwarf galaxies inside a halo. But observation say the opposite. An explanation can be that small halos aren't capable of keeping stars. We have also sub halos with dark matter only, sometimes called dark galaxies.

6.2 Problems with MOND

- Clusters and super-clusters: Although MOND can explain the missing mass in a galaxy, it can not explain the missing mass in a cluster, let alone a super-cluster. There is far too much acceleration in these two cases than MOND can predict.
- No relativistic version: you can see the Newtonian dynamics as an approximation of relativity theory, if you have low velocities and weak gravitational fields. If MOND is correct, you need a relativistic version, which is applicable for extreme cases of strong gravitational fields and velocities comparable with the speed of light. This relativistic version will of course be approximated by MOND in non extreme situations. As of yet, nobody has managed to construct a relativistic version of MOND.
- Gravitational lensing: Because of what we said above, MOND has no explanation of gravitational lensing phenomena, which is a completely relativistic effect.
- MOND cannot explain how CMB fluctuations lead to the present structure of the Universe

6.3 Scoreboard

Here we try to make a scoreboard for DM and MOND.

model / phenomena	DM	MOND
rotation curves of galaxies	+	+
Tully-Fisher relation	−	+
clusters and super-clusters	+	−
gravitational lensing	+	?
Evolution of the Universe from CMB	+	−

Table 1: scoreboard

One problem with MOND is it's isolation from the rest of physics. If MOND is correct, we must very likely change all of physics. IN other words, it is probably an approximation of other theories. (Look at section 6.2, No relativistic version.)

6.4 DM and MOND in the Solar System?

This section is maybe surprising. But scientists who have worked with pioneer 10 and 11 have discovered some acceleration towards the sun which is much too high to be from the solar system. They tried to find the explanation by looking for some source of error, but nothing seemed to work. Some people have actually tried to explain that by dark matter in the solar system or using MOND.

7 Conclusion

As you see, neither DM or MOND is currently adequate to explain the mystical phenomena referred to in the introduction. The majority of astronomers believe in DM, but as you

know we don't use voting to decide if a theory is right or wrong. The only thing we can use is hard evidence, which is not available right now. In the history of science and physics, people have been forced to abandon accepted theories and build completely new ones to explain new phenomena. Perhaps we are presently facing such a revolution.

More experiments and measurements are needed on this subject, and we need to be open-minded when it comes to new theories.

Acknowledgements

I want to thank Alessandro Romeo for all help and support he gave during the whole course and his ability to make physics a pleasure. I also want to thank my fellow students, together with whom we made the presentation possible. I want also thank Daniel Persson and Christoffer Petersson, who have let me study their work on the subject, and Peter Berntsen, who spend time to read my paper and correct my English.

References

Aguire A., 2003, arXiv:astro-ph/0310572
Anderson J.D., Laing P.A., Lau E.L., Liu A.S., et.al, 2002, Phys. Rev. D, 65, 082004
Binney J., Gerhard O., Silk J., 2001, Mon. Not. Roy. Astron. Soc. 321, 471
Binney J., 2003, arXiv:astro-ph/0310219
Ciotti L., Binney J., 2004, arXiv:astro-ph/0403020
Dunkel J., 2004, ApJ. 604, L37-L40
Gondolo P., 2004 arXiv:astro-ph/0403064
Jordan T.F., 2004, arXiv:astro-ph/0402384
Luscher R., 2003, arXiv:astro-ph/0305310
Milgrom, M., 2002, New Astron. Rev. 46, 741
Rees M.J., 2004, Phil. Trans. Roy. Soc. Lond. 361,2427
Sahni V., 2004, arXiv:astro-ph/0403324
Sanders R.H., McGaugh S.S., 2002, Annu. Rev. Astron. Astr. 40, 263
Sellwood J.A., 2004, arXiv:astro-ph/0401398
astron.berkeley.edu/ mwhite/darkmatter/rotcurve.html
http://blueox.uoregon.edu/ jimbrau/astr123/Notes/Chapter24.html
msowww.anu.edu.au/ pfrancis/astr1001/ notes/DarkMatter.PDF
www.astro.queensu.ca/ dursi/dm-tutorial/dm0.html
www.astro.umd.edu/ ssm/mond
www.astro.uu.se/ nisse/kurs/kos2002/DM.pdf
www.noao.edu/staff/shoko/tf.html
zebu.uoregon.edu/ imamura/209/apr7/TF.html

Probing the Acceleration of the Universe

Martin Nord

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(f00mano@dd.chalmers.se)

*

Abstract

Basic concepts and results in cosmology are introduced, and the expansion of the universe is discussed. On a basis of general relativity and the geometry of space-time, a more formal treatment of basic cosmology is also given, along with accounts of the fate of the universe in different cosmological models. Cosmological constraints from supernova observations, due to Perlmutter et al. (1999) and Knop et al. (2003) are presented and discussed. In particular, two recent results are accounted for; namely that the universe is accelerating and that dark energy is present. Finally, elementary results from other branches of observational astronomy are incorporated, and an emerging unitary picture of our universe i presented.

1 Introduction

Cosmology is the ancient field of science dealing with the grandest scales of the universe. It has long since been an integral part of metaphysics, but with the observational techniques of astronomy in modern times, cosmology has slowly but surely worked its way into physics.

Measuring quantities relevant to cosmology, however, is a precarious matter, for many physical properties of astronomical objects cannot be measured directly. When it comes to astrophysical obsevations, we can look, but we cannot touch. Physical distances, for example, have to be inferred from observable properties and are often inaccurate. Also, many different factors can contribute to the apparent brightness of a source, and it is not always easy to tell these factors apart. In a way, conducting observational astronomy is much like standing in a vast forest with a meter stick and the task of determining the heights of distant trees. The task seems easy enough until you are told you cannot move from the spot where you are standing. Indeed, this perspective is good to keep in mind when interpreting cosmological results.

In cosmology, the problems are far worse than those brought about by the difficulties in observational astronomy. If indeed the universe is infinite, there are no means for us to probe any significant part of it, and there never will be. We make observations of the

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

skies out to unfathomable distances, but these distances may well be vanishing in size when compared to the vast scales of cosmos. There is in effect no other way out than to make explicit assumptions about the nature of space and time outside our limited view of the universe.

Setting these difficulties aside, cosmology has during the last few years seen what is probably the most exciting period of discovery yet. Modern cosmology has developed during the 20th century owing much to the general theory of relativity put forth by Albert Einstein in 1916. The main problem of Newton's theory of gravitation – a problem which Newton himself was aware of – was the apparent "action at a distance" in his gravitational equation giving the gravitational force between two bodies of masses M and m separated by a distance r:

$$F = G\frac{Mm}{r^2} \tag{1}$$

(here G is the gravitational constant). This equation gives no information about how the two masses "know" how to, and in particular *when* to, fall toward each other. The general theory of relativity, by contrast, is a *local* theory of gravitation, asserting that mass (and other energy) curves space-time and thus causes massive objects to fall toward each other. A body cannot feel the mass of a distant massive object other than through the local curvature of space-time caused by that object, and the curvature is one of both space and time – there is no instantaneous interaction in space.

While general relativity successfully explains the kinematics of the sun and the planets, it can also be applied on cosmological scales. Einstein's main concern was that his equations did not have any solutions compatible with his notion of a static universe; gravity can only pull, not push, so an initially static universe would inevitably fall in on itself. To deal with this issue, he introduced into the equations a "cosmological term" with a fine-tuned negative pressure so as to exactly counterbalance the effect of gravity.

When it was discovered that distant galaxies are actually receding from us with velocities proportional to their distances, Einstein rejected the cosmological term, and it was thought to be gone once and for all. However, as we shall see, current research suggests that the cosmological term, often dubbed *dark energy*, is actually dominating the present expansion of the universe.

To obtain credible cosmological results, it is necessary to combine several different types of observations and see where they are consistent with each other. In recent years, a particularly promising method of determining cosmological parameters has been the study of distant supernova explosions. Such studies, combined with a host of other observations, are beginning to give a fairly consistent picture of the geometry and expansion of our universe.

This paper is written in such a way that a novice to astrophysics should be able to understand the basic aspects. Such a reader may, however, want to skip sections 3 and 4.3 in a first reading, as the rest of the paper can be understood independently of these in-depth sections.

2 Elements of Cosmology

An excellent starting point for cosmology is the simple observation that the night sky is dark. Consider an observer in a static universe, infinite in size and age and uniformly filled with luminous matter (stars). No matter in which direction the observer looks, his line of sight is eventually intercepted by a star, the result being a sky glowing with about

the same power as the surface of the sun. This argument, known as *Olbers' paradox*, provides a first clue that the universe, if infinite, is not static.

2.1 Hubble's Law

A definite solution to Olbers' problem was provided in the 1920's by Hubble and Humason, who, by studying frequency shifts of light from other galaxies, were able to conclude that the universe is expanding. The expansion stretches the light waves and shifts their energy (redshift), while at the same time it takes longer for a given photon to reach an observer. These factors are enough to account for the fact that the night sky is dark, even in a spatially infinite universe.

Distant objects, such as galaxies and clusters of galaxies, are moving away from us at speeds roughly proportional to their distances from us. The relation between recession velocity v and distance d is known as Hubble's law. It reads

$$v = H_0 d, \tag{2}$$

where H_0 is the *Hubble constant*, usually given in units of km s^{-1} Mpc^{-1}. A currently accepted value of this parameter is about 72 km s^{-1} Mpc^{-1} (Freedman et al. 2001). The relation holds regardless of the direction in which we look on the sky, but this does not mean that there is anything special about our part of the universe. The situation is quite similar to that of blowing up a balloon; an observer at any point on the surface of the balloon sees a recession of other points in any direction. Tracing the expansion backwards in time leads us to conclude that the universe was once extremely dense and hot. This, of course, is the basis of the well-known *big bang* model of the universe.

The expansion is often described in terms of the redshift suffered on the path through space. Let λ_e be the emitted wavelength from a source, and let λ_o be the observed wavelength here on earth. These wavelenghts are related by

$$1 + z \equiv \frac{\lambda_o}{\lambda_e}, \tag{3}$$

where the redshift parameter z is introduced as a convenient measure of cosmic distance: distant objects have large redshifts and large values of z. The cosmological redshift is often ascribed to the Doppler effect, the effect responsible for the fact that a fire truck sounds differently depending on whether it is moving towards you or away from you, but it is more correctly thought of as a stretch of the light wave due to the ongoing expansion of the universe, as illustrated in figure 1.

Similarly, though it may seem as though the distant galaxies are moving away from us *through space*, the galaxies are actually just moving along with the expansion of space itself – the whole cosmological coordinate system is expanding. Once again, think about the balloon. "Galaxies" drawn on the two-dimensional surface are not expanding into a surface that was not known before, but it is the whole surface that is stretching. As a consequence of this, we can define *comoving coordinates* such that the spatial coordinates of a galaxy at rest with respect to the cosmic background radiation (i.e. a galaxy following the mean expansion of the universe) are constant. In the balloon example, two points on the surface will maintain the two angular coordinates as the balloon inflates (figure 2). The *radius* of the balloon will change, but this coordinate is not relevant to local distances on the two-dimensional surface (it is, however, relevant to cosmological observations as we shall see later). Also, the *physical distance* between the points will of course change.

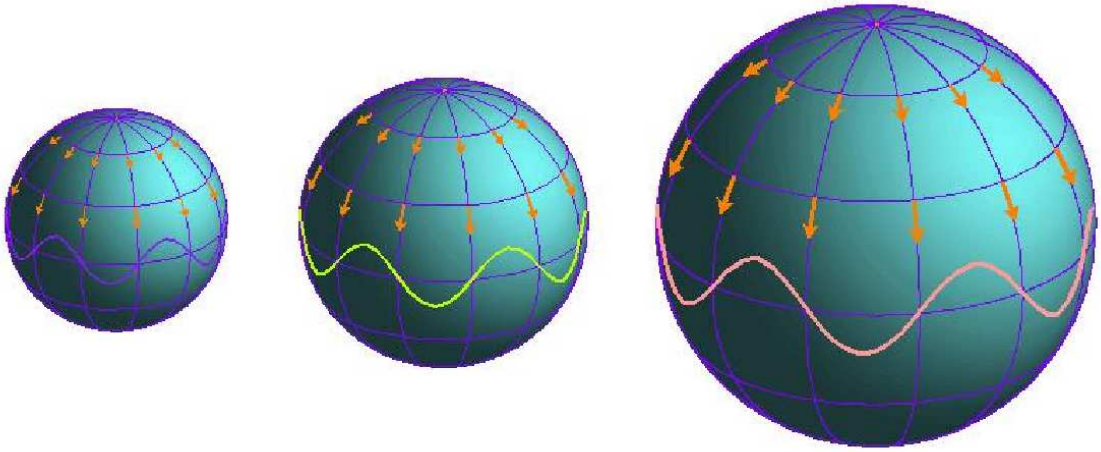

Figure 1: Two-dimensional analogy to the expanding universe. On an expanding sphere, there is nothing special about a point from which one sees equal recession speeds in all directions. Furthermore, the redshift of a light wave is just a stretching that follows the stretching of space itself. Adapted from Wayne Hu's Introduction to the Cosmic Microwave Background: `http://background.uchicago.edu/ whu/beginners/introduction.html`

In three-dimensional space, by analogy, we have three comoving spatial coordinates as the physical distance between any two comoving points increases. The three-dimensional analogy to the radius of the balloon is called the *scale factor* $a(t)$ of the universe. The t appearing here is the *cosmic time* variable, which is most simply thought of as the time elapsed since the dense, early period of the universe (when there was no spatial separation to speak of) measured at any comoving point in space.

2.2 The Cosmological Principle

With the comoving coordinates of cosmos in place, we can state a very important principle on which modern cosmology rests. Given that the universe is expanding, it is quite certain that it has a finite age. This, in turn, means that there is a limit to how far out in the universe we can look, set by the fact that light cannot have traveled for a time longer than the age of the universe. The surface out to which we – in principle – can observe is called the *horizon*. Although the horizon is steadily growing larger as light from the early universe is given more time to reach us, we will never be able to see arbitrarily far in an infinite universe.

Knowing nothing about the structure of the universe outside our cosmic field of view, we must make assumptions about space and time outside the horizon. The most natural assumption, of course, is that the distant universe looks the same as our neighborhood of galaxies and galaxy clusters. This is not merely a matter of convenience; compelling theoretical arguments for such uniformity are given by Weinberg (1972) and by Zwicky and Rudnicki (1963). In particular, we state the assumption known as the *cosmological principle*: On large enough scales, the universe is *isotropic* (it looks the same in all directions) and *homogeneous* (it looks the same at any given point). A homogeneous space is not necessarily isotropic, but a space that is isotropic about every point must also be homogeneous. It is the latter that is thought to be the case with our universe.

The cosmological principle is very useful when dealing with the geometry of space-

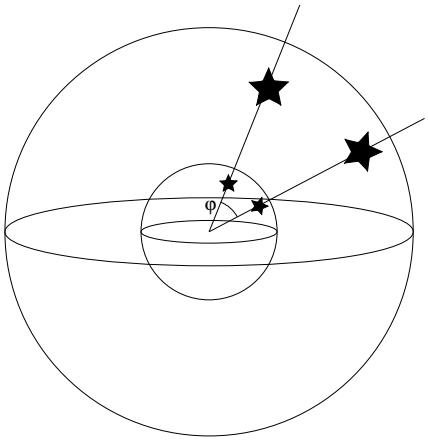

Figure 2: Comoving coordinates in a two-dimensional spherical space. The angular coordinate φ between two points following the expansion of the space remains the same, while the physical distance between the points is stretched.

time, as it provides a simple way for us to extrapolate the structure of the universe outside our very limited view of it. This, in turn, gives us a way of dealing with the curvature of space-time on cosmological scales.

2.3 General Relativity and the Geometry of the Universe

From childhood, the Euclidian concepts of geometry are deeply rooted in our minds, but there are simple instances in which Euclidian geometry does not hold. On the surface of a balloon, for example, the angles of a triangle always add up to more than 180 degrees, and two initially parallel lines eventually intersect. Put simply, the balloon surface is not flat; it possesses what is known as *positive curvature*. In the same way, space-time is not necessarily flat, at least not everywhere. This is harder to picture, but it is simply a four-dimensional analogy to the balloon scenario. Thus, space-time can possess positive curvature, or initially parallel lines might eventually diverge, in which case we speak of *negative curvature*. In the case of parallell lines remaining parallell, the geometry is said to be flat. It should be noted that space-time, unlike the balloon, is not necessarily curved *into another dimension*. By analogy, a two-dimensional surface with constant negative curvature is perfectly feasible, though it cannot be embedded in three-dimensional Euclidian space.

Curvature is dealt with through the *metric function*, which relates the coordinates of two nearby points in space-time to the distance between those points. In flat euclidian space, the metric function simply reduces to Pythagoras' theorem. In flat space-time, the metric is the *Minkowski metric* of special relativity:

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2. \tag{4}$$

Here, the line element ds^2, also known as *proper time*, is independent on the inertial frame of reference. Note that this also looks like Pythagoras' theorem apart from the difference in sign between time and space; the Minkowski metric does in fact describe a flat space-time.

In general, the metric function of space-time is not as simple as the Minkowski metric. The key concept of general relativity is that energy, such as mass, curves space-time, and it is therefore the abundance of energy and matter that determines the curvature.

Curvature, in turn, is the cause of gravity – a massive object curves space-time around it in such a way that the paths of nearby objects or particles are bent toward it (Note that the paths of massless particles, such as photons, are also bent). In the succinct wording of Charles W. Misner, "space acts on matter, telling it how to move. In turn, matter reacts back on space telling it how to curve" (Misner et al. 2002).

At a glance, it might seem as though general relativity, with its geometric interpretation of gravity, introduces unnecessary complications into the physics of gravitation. When studied in detail, however, it is seen that general relativity is actually quite a natural approach. Consider two hikers standing 100 meters apart on the equator of the earth, and then starting to walk due north. Initially their paths will be parallell, but when they reach the north pole the paths will intersect. There are two ways in which we may explain the situation. We can state that the paths meet because the earth has positive curvature, and so parallell lines eventually converge, or, conversely, we may introduce a force between the hikers and call it "gravity". Undoubtedly, the first alternative is the most compelling.

Formally, the general theory of relativity amounts to *Einstein's field equations*:

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{5}$$

Here, $R_{\mu\nu}$ is the Ricci tensor (describing the curvature of space-time), $g_{\mu\nu}$ is the metric tensor and G is the gravitational constant. The terms on the left are concerned with geometry, while the right-hand side is the source term; $T_{\mu\nu}$ is the so-called energy-momentum tensor, containing all forms of energy. The non-linearity of these equations – arising from the fact that the gravitational field interacts with itself – makes them very difficult to solve, and a general solution does not exist.

The cosmological principle gives the metric of the universe as a whole, and substituting this into (5) gives the cosmological equations describing the kinematics of the universe. These equations are not consistent with a static universe, however, and for this reason Einstein added a term $-\Lambda g_{\mu\nu}$ to the left-hand side of the equations (alternatively, it could be included in the source term). Λ is the famous cosmological constant, which, when fine-tuned, makes a static solution possible. Today we know that the universe is expanding, but recent findings point to the existence of a non-zero Λ, arising from an exotic entity called dark energy.

2.4 Dark Matter and Dark Energy

It has long been known that there is "missing" matter in the universe. The dynamics of typical disk galaxies are generally not compatible with the observed abundances of luminous (baryonic) matter, and kinematics within galaxy clusters also suggest there is more mass than we can see. What this so-called *dark matter* consists of is not settled, but recent developments suggest it is largely non-baryonic in nature (Sahni 2004), and that it is constitutes at least 90% of all matter in the universe.

Dark matter is highly relevant to cosmology as the matter density of the universe plays a crucial role in determining its fate. A universe with almost no matter will expand forever, but if the density is high enough, gravity can eventually halt the expansion and bring about a gravitational collapse. The *critical density* of the universe, i.e. the minimum density required for gravity to reverse the cosmic expansion, is

$$\rho_{crit} = \frac{3H^2}{8\pi G}. \tag{6}$$

In the present context, the Hubble "constant" H is dependent on cosmic time (H_0 is the value of H today). If H changes over time, then the critical density also changes since it takes different amounts of gravitational energy to counteract different values of kinematic energy of expansion. In section 4 we shall explore the time dependence of the Hubble constant, which simply amounts to an acceleration or deceleration of the cosmic expansion.

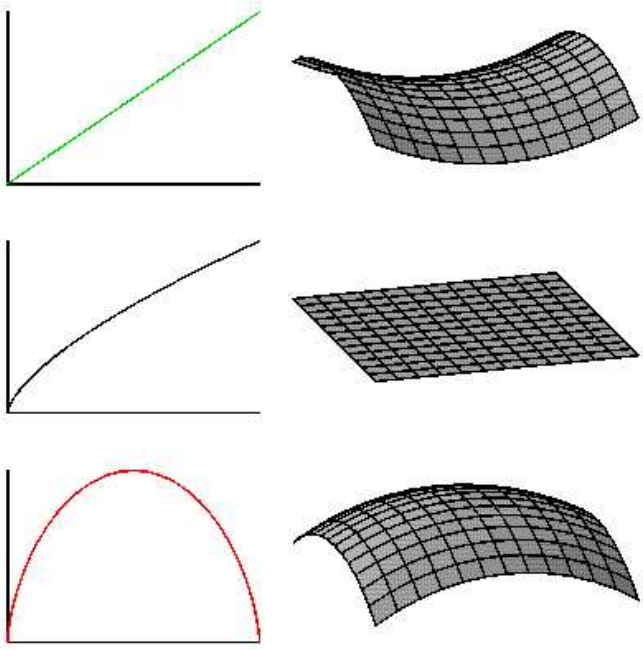

Figure 3: Three curvature cases (2-dimensional analogy) shown along with corresponding time evolution of the scale factor in an entirely matter-dominated universe. Graphs show $a(t)$ versus cosmic time. From Ned Wright's Cosmology Tutorial: `http://www.astro.ucla.edu/ wright/cosmolog.htm`

The ratio of gravitational energy to kinetic energy from the cosmic expansion is described in terms of the variable Ω, which is just a density measure in units of the critical density. If Ω equals one, then there is perfect balance between gravity and the energy contained in the expansion; thus the value never changes. This is the case of a flat universe. If, on the other hand, Ω is different from one, the value increases or decreases very rapidly, giving rise to either an out-of-control expansion or a quick demise through gravitational collapse.

As stated before, curvature and energy content in the universe are not independent, and the simple relation is that $\Omega = 1$ corresponds to a flat universe, while $\Omega > 1$ and $\Omega < 1$ correspond to a closed and an open universe, respectively. Figure 3 shows the three curvature cases (open, flat and closed universe) with the corresponding plots of $a(t)$. The curves assume, however, that the cosmological constant is zero, which is not the current standard model. Including Einstein's cosmological term into the field equations will drastically change the situation. The Λ term, if positive, represents a form of energy which has negative pressure and so acts as a kind of "anti-gravity," contributing to the cosmic expansion rather than slowing it down. It is precisely the rate of change of the Hubble expansion that determines whether or not dark energy needs to be included in the cosmic picture, although it has been difficult to find a credible physical explanation for Λ (Cohn 1998).

Figure 4 shows how the introduction of dark energy changes the fate of the universe. Even for a universe with a matter density higher than the critical one, the universe is

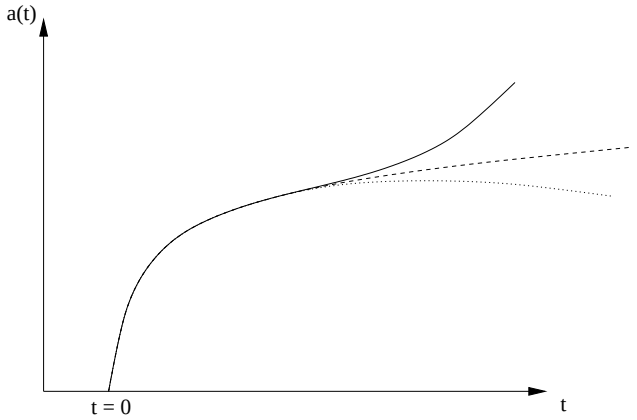

Figure 4: Schematic view of the fate of our universe, with cosmological length scale plotted against cosmological time. Solid line: The expansion is slowed as long as the matter density is high enough to dominate over dark energy. The universe becomes increasingly dilute, and dark energy eventually takes over, causing the expansion to accelerate. Dashed line: No dark energy is present, but there is not enough mass to completely halt the expansion. Dotted line: Matter-dominated universe in which the density is high enough to recollapse the universe.

not destined to re-collapse. The definition of the critical density in (6) thus only gives the matter content necessary to re-collapse a universe entirely composed of matter, but it nonetheless gives us a starting point for cosmological survey. Observations suggest that the matter density of the universe is less than, but comparable to, the critical density (Krauss 1999). Including dark energy as a contribution to the energy density of the universe still means the universe is flat for $\Omega = 1$, where Ω is for all types of energy.

Today, the energy density in the universe is neither infinite nor zero (obviously). In fact, the present rate of cosmic expansion is very close to making the universe flat. This means that Ω must be equal to, or *very* close to, one. Given the estimated age of the universe (at least ten billion years), this seems like quite unnatural a coincidence (see Hogan 1999).

We can separate the contributions to Ω from different types of energy. The contribution from matter, Ω_M, is approximately 0.3, of which only a small fraction is from baryonic matter (see, for example, Turner 2002). Dark energy has negative pressure but positive energy and so contributes to Ω if present. We can also have a contribution from the curvature of space-time (see below), but observations of the cosmic microwave background (CMB) radiation suggest the universe is flat or very close to flat (Ruhl et al. 2003). If the latter is correct, it is required that dark energy constitutes a significant contribution to Ω.

There is, however, an important argument against the cosmological term: If Λ is constant, then why do we happen to live at a time when the dark energy density is comparable to that of matter? A novel approach, Quintessence, states that the cosmological constant is in fact not constant at all. We need not be concerned with such issues here, but the interested reader is encouraged to study Daniel Johansson's article.

3 Formal Development of Cosmology

This section is meant to give a deeper understanding of the physics of cosmology. Parts of the previous section are studied in greater detail, and are given a more formal treatment. Though such a treatment gives a fuller insight into the marvels of the universe, it is not necessary for understanding the cosmological experiments and results presented of the next section. The account given here is by no means a complete one, but merely a collection of theoretical results relevant to this paper. For a more detailed account, see Longair (1998).

3.1 The Metric

The cosmological principle discussed in the previous section sets some fundamental limitations on the large-scale metric of space-time. A space that is both isotropic and homogeneous is maximally symmetric. Furthermore, maximally symmetric spaces with a given curvature constant are all equivalent, so if we can construct one such space, we will know it is the right one. A construction method is described by Bergström and Goobar (1999), and the resulting metric is the Friedmann-Robertson-Walker metric

$$ds^2 = dt^2 - a^2(t)(\frac{dr^2}{1-kr^2} + r^2 d\theta^2 + r^2 sin^2\theta d\phi^2), \tag{7}$$

where r, θ and ϕ are spherical polar coordinates and $a(t)$ is the cosmic scale factor, describing how distances in the universe scale with time. k is a curvature parameter that can only assume the values 0 (flat space), $+1$ (positive curvature) or -1 (negative curvature). In fact, k is just the sign of the curvature scalar of general relativity, while the length scale of the (spatial) curvature is absorbed into $a(t)$.

The most important point to make about this metric is that it is just a result of implementing the cosmological principle, not of solving Einstein's field equations. On the contrary, we can now use the field equations and the metric to derive the equations of motion for the universe.

3.2 Solving Einstein's Equations

We shall be concerned with solving the field equations involving a cosmological constant Λ:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{8}$$

The energy-momentum tensor is taken to be that of a perfect fluid, namely

$$T_{\mu\nu} = (p+\rho)u_\mu u_\nu - p g_{\mu\nu}, \tag{9}$$

where p and ρ are the pressure and energy density of the fluid. u_μ is the four-velocity, which in comoving coordinates only has a time-component; $u_\mu = (1,0,0,0)$. The equation of state is simply

$$p = w\rho, \tag{10}$$

where the parameter w takes the values $w_\Lambda = -1$ for dark energy, $w_{rad} = \frac{1}{3}$ for radiation and $w_M = 0$ for non-relativistic cold dark matter (which is currently believed to be the main constituent of matter in the universe). At any time, the universe contains a mixture

of these three types of energy, so that $\rho = \rho_\Lambda + \rho_m + \rho_{rad}$ and $p = p_\Lambda + p_{rad}$. Taking the tt- and rr-components of (8) with $T_{\mu\nu}$ from (9) gives, after some algebra,

$$\left(\frac{\dot{a}}{a}\right)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\rho; \tag{11}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p), \tag{12}$$

where the dark energy contribution to ρ has been taken as $\rho_\Lambda = \frac{\Lambda}{8\pi G}$, and dots denote time derivatives. Equations (11) and (12) are known as the Friedman equation and the acceleration equation, respectively. The first term in (11) relates velocity to comoving distance and is simply the Hubble parameter $H(t)$ sqared. It is immediately seen from (12) that a universe with $\rho + 3p < 0$ has a positive acceleration of the expansion. Energy density is always positive, so the only thing that can cause an accelerating expansion is dark energy with negative pressure.

From the conservation of energy-momentum, it is straightforward to derive another useful equation. Setting $\nabla_\mu T^{\mu\nu} = 0$, one finds

$$\frac{d}{dt}(\rho a^3) = -p\frac{d}{dt}a^3. \tag{13}$$

Note that a^3 is just a volume, so the left-hand side of (13) is a time derivative of energy. Written in another way, $dE + pdV = 0$, which just expresses that there is no change in entropy for a perfect fluid.

3.3 Cosmological Solutions

Measuring the age of the universe requires relating $a(t)$ to cosmic time t. To do this, we first have to derive expressions for ρ and p. Substituting the equation of state into (13) and solving the resulting differential equation yields

$$\rho \sim a^{-3(1+w)}. \tag{14}$$

In a matter-dominated universe, where $w = 0$, this means that $\rho_m \sim a^{-3}$, and in a radiation-dominated universe, with $w = 1/3$, $\rho_{rad} \sim a^{-4}$ (the extra factor of a^{-1} is readily interpreted as energy loss due to the red shift of radiation).

It is now straightforward to solve the acceleration equation. Using (14), we find

$$a(t) \sim t^{\frac{2}{3(1+w)}}. \tag{15}$$

Thus, for matter $a(t) \sim t^{-2}$ and for radiation $a(t) \sim t^{\frac{1}{2}}$.

Of course, the universe is neither radiation-dominated nor matter-dominated throughout it's entire lifetime. Ignoring dark energy it is today dominated by matter, but when it was younger than about 350 000 years, in the period before recombination, it is believed to have been radiation dominated. We thus have to solve for $a(t)$ in a variety of cases, varying the dependence on k, $\rho(t)$, $p(t)$, and including Λ.

To solve the Friedman equation, we simply divide by $H^2(t)$ to find

$$1 + \frac{k}{H^2a^2} = \frac{8\pi G}{3H^2}\rho. \tag{16}$$

Defining ρ_{crit} as in (6) yields

$$1 + \frac{k}{H^2 a^2} = \frac{\rho}{\rho_{crit}} \equiv \Omega(t). \tag{17}$$

The last identity is the definition of the density parameter Ω. For any value of t we see that if $\Omega > 1$, then $k = 1$ and if $\Omega < 1$ then $k = -1$. A flat universe occurs only for Ω exactly equal to one. Depending on the contributions to Ω from different types of energy, equation (17) will behave differently as t goes to zero. It is noteworthy, however, that Ω approaches 1 as $t \to 0$.

The importance of equation (17) is that it gives us the coupling of energy density and curvature. A closed, matter-dominated universe with no cosmological constant will inevitably recollapse as it requires $\Omega_M > 1$ (In popular literature it is often stated that a closed universe will eventually recollapse, while an open universe with negative curvature will expand forever. This is not strictly true, since we must take into account not only matter but also dark energy). As mentioned earlier, CMB experiments point to a flat universe, i.e. $k = 0$, but there is clearly not enough matter present to make it so. Currently, this is the one of the main pieces of observational evidence for the presence of dark energy.

To see how the density parameter depends on different forms of energy, we define

$$\Omega_M = \frac{8\pi G}{3H^2}\rho_m, \tag{18}$$

$$\Omega_\Lambda = \frac{\Lambda}{3H^2}. \tag{19}$$

Equation (17) now reads

$$1 = \frac{-k}{a^2 H^2} + \Omega_M + \Omega_\Lambda. \tag{20}$$

(In some literature, the first term in equation (20) is called Ω_k. We shall not follow this custom here, as this is not an energy density contribution in the ordinary sense.)

3.4 The Fate of the Universe

To deal with the ultimate fate of the universe as we know it, we have to consider a variety of models with different values of Λ and different values of k. For a complete treatment we would have to take into account nine distinct cases; no Λ, positive Λ and negative Λ, each with three different possible values of k. However, for reasons mentioned above, we shall assume Λ is non-zero and positive. The $\Lambda = 0$ case is described in a plethora of literature; for an excellent introduction see Longair (1998). As we live in a universe in which the energy density of radiation can be neglected as $t \to \infty$, we shall only be concerned with ρ_m and ρ_Λ in the relevant equations.

The Friedmann equation becomes more transparent if we separate the impact of matter from that of dark energy. Noting that $\rho = \rho_0(a_0^3/a^3)$, where the naughts in the subscripts indicate present parameter values, equation (11) is readily rewritten as

$$\dot{a}^2 = \frac{8\pi G}{3}\frac{\rho_0 a_0^3}{a} - k + \frac{\Lambda a^2}{3}. \tag{21}$$

Note that in this equation, ρ pertains only to the energy density of matter. It is immediately seen that the Λ term can be neglected as $a \to 0$. Conversely, for large values of a,

dark matter becomes dominant. For $k = 0$ and $k = -1$, (21) gives $\dot{a} > 0$, which means the universe expands forever. With $k = +1$, however, it is possible to fine-tune Λ in such a way that $\dot{a} = \ddot{a} = 0$.

In the present section, we have seen the deep interconnections between the cosmological parameters. In the following, we shall have a look at one particular way of probing some of these parameters: the study of high-redshift type Ia supernovae.

4 Cosmological Constraints from Supernova Observations

It has long since been expected that galaxies far away from us might depart from the linear Hubble expansion given in equation (2); either because the rate of expansion is changing or because the intervening space is not flat. Probing the curvature and expansion rate of the universe, however, is by no means an easy task. Because the speed of light is finite, looking out in the universe means also looking back in time. Observing objects at high redshifts involves large uncertainties, particularly because we do not know for sure how the universe was composed at earlier epochs. Furthermore, it is not trivial to measure luminosity (power) and distance independently of each other, as the Hubble constant is not known a priori. The ideal solution would be placing standardized light bulbs with known power output all over the universe, and then determining distances from the relation

$$F = \frac{L}{4\pi d^2}, \tag{22}$$

where L is the luminosity of the light bulb and F is the apparent brightness, measured here on earth in units of energy per unit of area and time. Of course, putting such light bulbs in place would prove quite a difficult task (to say the least), but what if they were already in place? The well-accepted models for stellar evolution suggest that one such "standard candle" is realized in the violent outburst of energy that occurs when a star collapses – a supernova explosion.

4.1 Standard Candles of the Universe

A supernova is the blast occurring from a dying star that can longer support the inward pull of it's own gravity. There are different types of supernovae, but we shall here be concerned with a subclass known as Ia. These supernovae occur in some binary star systems consisting of a red giant and a white dwarf. In such a system, mass flows from the red giant to the white dwarf, and eventually so much mass piles up on the white dwarf that it can no longer support itself. The star then implodes, but the implosion can usually be brought to a halt by neutrons, the only things in nature that can stop such a gravitational collapse (even neutrons sometimes fail depending on the mass of the star's core). When the collapse is abruptly stopped by the neutrons, matter bounces off the hard core, thus turning the implosion into an explosion. It is the light from such an implosion that is thought to form a cosmic "standard candle".

There are slight differences in the power of the type Ia supernovae, but there is a way of correcting for this: brighter explosions last longer than fainter ones. By taking this into account, it is possible to deduce the inherent brightness of a supernova to about 12% (Branch 1998).

A supernova explosion can easily outshine a whole galaxy, but it is quite a rare event; in a typical galaxy, a type Ia supernova is expected to occur about once every 300 years. There are, however, many galaxies in the universe, and finding distant supernovae is just a matter of taking images of the sky a few weeks apart and comparing them to see if there are any significant changes that could be exploding stars.

4.2 Probing Space-Time

Already initial studies indicated that the distant supernovae were up to 25% dimmer than what one would expect from a flat universe with a linear Hubble law (Hogan 1999). One could attempt to explain this dimming with cosmic dust spread through space, but dust grains would filter out blue light more than red, and no such effect is seen in observations. How then, shall we interpret such results?

Firstly, it may be the case that space has negative curvature. This would mean that the light from a supernova would spread over a larger volume than in flat space, causing it to appear dimmer when observed over great distances. To see this, the balloon analogy once again comes in handy. Recall that the balloon has positive curvature, and that a circle of radius r has a circumference which is smaller than 2π. Now imagine a supernova explosion in the center of the circle, observed by an astronomer somewhere on the perimeter. All light from the supernova will eventually have to pass the perimeter of the circle, but as it is smaller than a corresponding circle in flat space, the density of the radiation will be higher, and the supernova will appear brighter than it really is. In the opposite case of negative curvature, the supernova would instead appear dimmer.

Secondly, the supernovae could be farther away than their redshifts from the linear Hubble law suggest (or, from a different point of view, they have less redshift than anticipated). This would mean that the universe expanded slower in the past (figure 5).

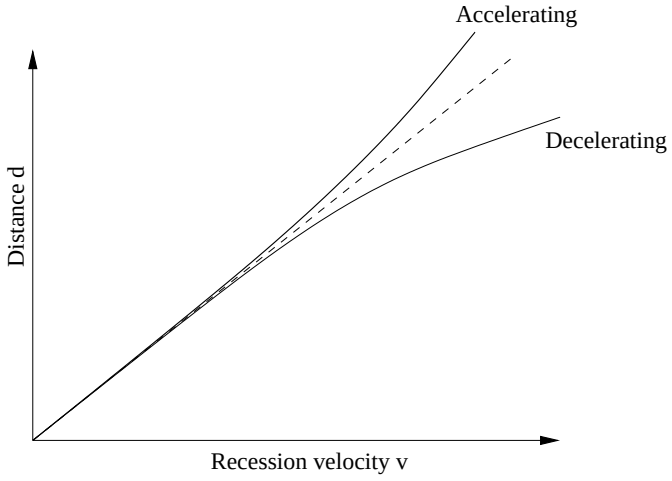

Figure 5: Deviation from the linear Hubble law. The thick solid line shows the distance-velocity relation which should be observed in an accelerating universe. Note that the quantities plotted here are not directly observable properties in astronomy.

The case for the accelerating universe is strong even with just these elementary pieces of information at hand. As we have seen, a small density of matter implies negative curvature as well as little gravitational slowing of the cosmic expansion. However, the observed supernovae are fainter than predicted even for a nearly empty universe, which implies that acceleration has to be at least part of the truth. Moreover, an accelerating

universe solves one important cosmological problem that was the source of much dismay during the 1990's: that of the age of the universe. Estimating the age of the universe using the linear Hubble law, one finds that some of the oldest stars in our Milky Way appear older than the universe itself. This is clearly impossible, but if the universe expanded slower in the past, the age of the universe has to be revised upward (for an in-depth account, see Krauss 1998).

With these basic arguments at hand, we shall now have a closer look at some recent observational evidence for the accelerating universe.

4.3 Method

Before plunging into the results, it is perhaps instructive to give an outline of the method used in constraining cosmological parameters from supernova experiments. The method accounted for here is essentially that described by Perlmutter et al. (1999) (hereafter P99), and used by Knop et al. (2003) (hereafter K03).

Measuring the rate of change of the Hubble parameter requires carrying out observations over large distances, i.e. the observed supernovae must be situated at large redshifts. P99 reports measurements of supernovae at $z = 0.18$-0.83 recorded using the Hubble Space Telescope (HST), and followed over the peak of their light curves for about 2-3 months (corresponding to 40-60 days in the rest frames of the supernovae). The supernovae used in K03 are at the same order of redshifts.

To standardize the measured luminosities, filters that correspond to rest-frame B and V filters are used, and any remaining wavelength mismatch is corrected for using a so-called "cross-filter K-correction", calculated based on template spectra from low-redshift supernovae. The light curves are furthermore normalized using a simple linear relation between peak luminosity and time axis stretch of the light curve shape.

After correcting for galactic extinction, one arrives at a final form for the relation between magnitude and redshift, as given in Perlmutter et al. (1997):

$$m_B^{eff} \equiv m_R + \alpha(s-1) - K_{BR} - A_R = \mathcal{M}_B + 5\log\mathcal{D}_L(z;\Omega_M,\Omega_\Lambda). \tag{23}$$

Here, K_{BR} and A_R are the cross-filter K-correction and the correction for galactic extinction, m_R is the observed R-band magnitude, and the parameters α and s are concerned with light curve stretch (see Perlmutter 1999). The luminosity distance d_L is given implicitly in $\mathcal{D}_L$ by $\mathcal{D}_L \equiv H_0 d_L$, and $\mathcal{M}_B$ is defined by $\mathcal{M}_B \equiv M_B - 5\log H_0 + 25$, where M_B is the absolute B-band magnitude at maximum of a Ia supernova with $s = 1$. The quantity m_B^{eff}, called the effective B-magnitude, is to be understood as a measure of standard luminosity at a given redshift.

This form of the magnitude-redshift relation is convenient because the quantity $\mathcal{D}_L$ is determined from theory independently of H_0. The "Hubble constant-free" B-band magnitude $\mathcal{M}_B$, fitted from apparent magnitudes, is also independent of H_0, and the cosmological parameters Ω_M and Ω_Λ can thus be determined without knowledge of the exact value of the Hubble parameter.

4.4 Results

The exact method by which supernova observations are fitted to the cosmological parameters is a complex procedure beyond the scope of this text, but the standard form of presenting the results is quite easy to understand given the concepts introduced in section 2. The *Hubble diagram* displays the relation between redshift and effective magnitude (or

a corresponding quantity). In a flat universe where $H = H_0$ at all redshifts, the effective luminosity of a standard supernova is just a linear function of redshift (magnitude is not a linear function of redshift, however, as magnitude is defined on a logarithmic scale), and the slope is readily interpreted as the Hubble constant.

The cosmological influence of parameters such as Ω_M and Ω_Λ is usually visualized in a plot of *parameter space*. Strictly speaking, this is a space containing as its dimensions *all* independent cosmological parameters, though only a few can be displayed at a time. Various cosmological experiments give constraints in various parts of parameter space. In the present context, we are only concerned with constraints in the subspace containing Ω_M and Ω_Λ.

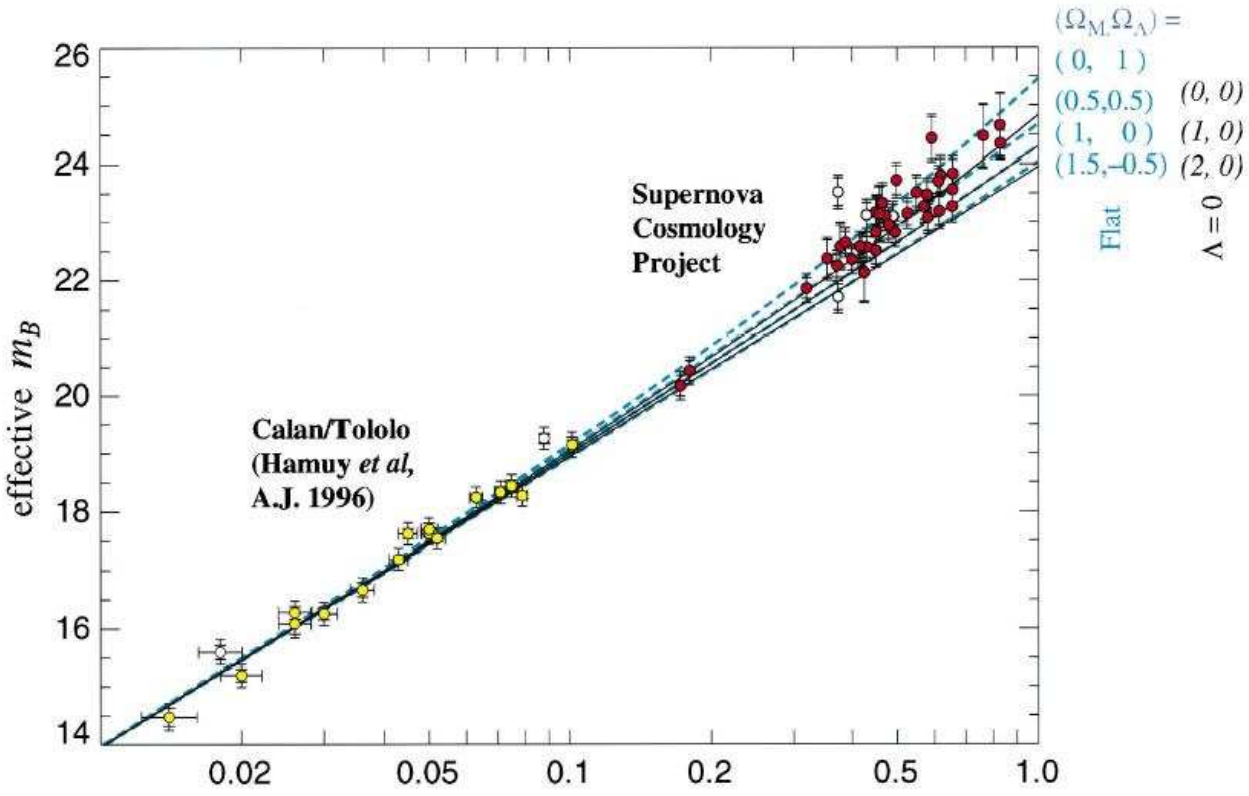

Figure 6: Hubble diagram for 42 high redshift supernovae from the Supernova Cosmology Project and 18 supernovae from the Calan/Tololo survey, from P99. Outer error bars show total uncertainties and unfilled circles indicate supernovae not included in the primary cosmological fit of P99. Solid curves are theoretical magnitudes for a range of cosmologies with zero Λ, and dashed curves are for a range of flat cosmologies with non-zero cosmological constant.

Figure 6 displays a Hubble diagram of 42 high redshift supernovae from an early stage of the *Supernova Cosmology Project*. The distant supernovae clearly deviate from the linear Hubble law (there is no way to fit a straight line to the results), and the theoretical curves given in the figure might offer a clue as to the reasons for this. Either the universe is flat and there is dark energy present to cause an acceleration, or space-time has negative curvature (or both). At this point there is no obvious way to discriminate between the two possibilities. Implementing the cosmological equations outlined above, however, enables us to go one step further and set limits on the energy content of the universe.

The primary fit to cosmological parameters from K03 is shown in figure 7, along with implications of the combination of the parameters on the fate and geometry of the

universe. Even the 99% confidence region indicates strongly that the universe is indeed

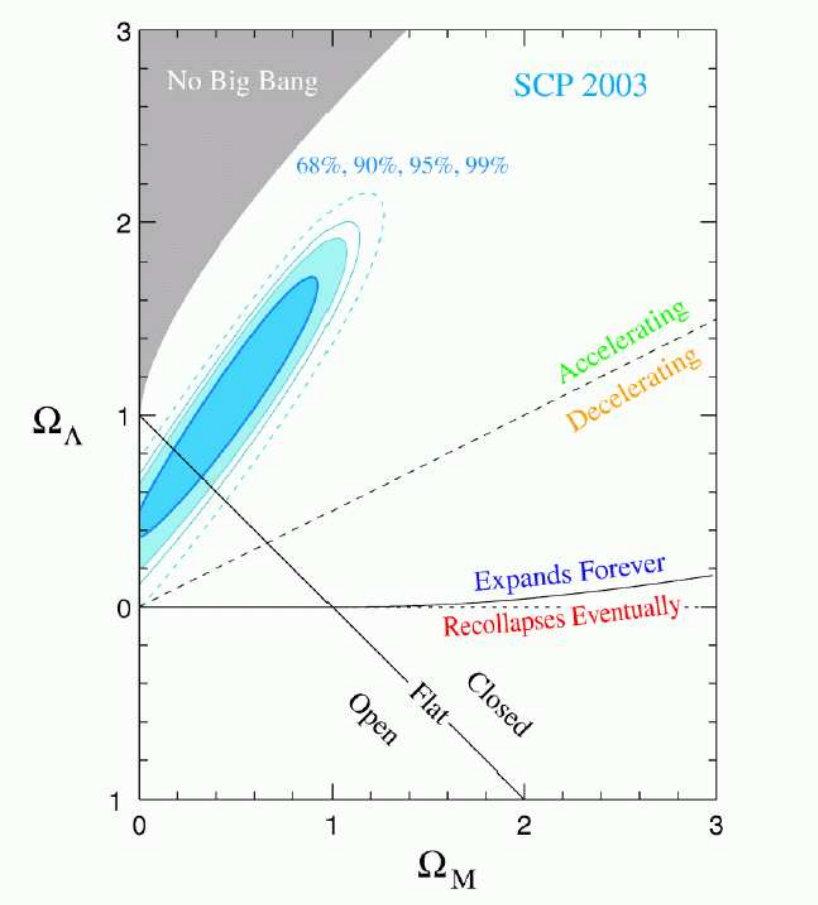

Figure 7: Confidence regions for Ω_M and Ω_Λ from the primary cosmological fit of K03. Note that the spatial curvature of the universe is not determinative of the future of the cosmological expansion, indicated by the near-horizontal solid line. The shaded region at the upper left represents a universe with a dark energy-contribution large enough to exclude a big bang, while the region at the lower right corresponds to a universe younger than the oldest heavy elements (see Schramm 1990).

accelerating and has a significant dark energy content. The results are clearly consistent with a flat universe, and moreover it seems inevitable from these observations that the universe will continue expanding forever.

5 Concluding Remarks

We have seen that the evidence for an accelerating universe is strong indeed, but we must carefully examine the results to make sure there is no alternative interpretation to the raw data. Firstly, it may be argued that high redshift supernovae could be different in composition from nearby ones, owing to the lower abundances of heavy elements in younger galaxies. Although this possibility cannot be excluded completely, it should be possible to take such effects into account; theoretical models indicate that differences in abundances of heavy elements should cause noticeable differences in spectral features (Höflich et al. 1998). Other possible "loopholes", such as Malmquist bias (the tendency

to sample objects with higher luminosity than the average) are discussed in P99, and the error budgets of P99 and K03 are adjusted accordingly.

The results in parameter space are clearly inconsistent with a flat $\Lambda = 0$ universe. The universe may be flat *or* there may be little or no dark energy, but the two are mutually exclusive. Combining the data with constraints from CMB observations, indicating a flat universe, points to the former alternative with a significant dark energy content.

To confirm the results mentioned above, there are many methods that have recently been suggested, i.e. probing space-time using gamma ray bursts (Girolamo et al. 2004). More interesting, however, are observational methods giving results orthogonal in parameter space to those of supernova observations. The CMB has already been mentioned, and more precise measurements of the anisotropies in the background radiation are to be expected within the next decade. Another promising method is the study of scattering of the the CMB photons from hot gas in galactic clusters (Sunyaev-Ze'ldowich effect), which provides an independent way of mesuring the Hubble parameter (see Battye & Weller 2003). Further supernova observations from the SNAP project (Sahni 2004) are also expected to give better constraints on Ω_Λ and Ω_M.

As results from a wide variety of observational methods are coming into increasing agreement, our current understanding of the universe can be summarized by the following:

- Space-time is flat or close to flat (from CMB anisotropies)
- The universe is accelerating
- The main constituent of matter is non-baryonic dark matter
- Dark energy is present, with a contribution of about 0.7 to Ω
- The temperature is roughly that of the background radiation, $T_r = 2.7$ K
- The value of the Hubble constant today is $H_0 = 72 \pm 8$ km s^{-1} Mpc^{-1}

In the theoretical development of cosmology, different modifications of the models described here are frequently suggested in order to accommodate theoretically favorable cosmologies. Freese & Lewis (2002) thus arrive at a flat, accelerating and matter dominated universe with no dark energy contribution by modifying the Friedman equation, while Canfora & Troisi (2003) propose an inhomogeneous model to resolve the "fine-tuning problem" of dark energy (for a popular introduction to this interesting problem, see Krauss 1999). Future experiments will tell whether such proposals will indeed prove useful, but as long as observational results from a wide variety of methods are in agreement, there is no immediate need to question the accuracy of the standard cosmological model presented here.

Acknowledgements

I would like to thank Daniel Johansson for interesting converstaions on the meaning of cosmology, and Stephanie Cline for love, support and proofreading the final manuscript. Special thanks also to Gösta Lindblad for offering insightful philosophical comments on the subject, and to Cathy Horellou at Onsala Space Observatory for answering more technical questions.

References

Battye R. A., Weller J., 2003, Phys. Rev. D68, 083506

Bergström L., Goobar A., 1999, Cosmology and Particle Astrophysics. Wiley, Chichester

Branch D., 1998, ARA&A 36, 17

Canfora F., Troisi A., 2003, Gen. Rel. Grav. 36, 273-385

Cohn J. D., 1998, ApJS 259, 213

Freedman W. L., 2001, ApJ 553, 47

Freese K., Lewis M., Phys. Lett. B540, 1-8

Girolamo T., Vietri M., Sciascio G., 2004, arXiv:astro-ph/0401244

Höflich, et al., 1998, ApJ 495, 617

Hogan C. J., Kirshner R. P., Suntzeff N. B., 1999, Scientific American 280, 46

Knop R. A., et al., 2003, arXiv:astro-ph/0309368

Krauss L. M., 1998, ApJ 501, 461-466

Krauss L. M., 1999, Scientific American 280, 52

Longair M. S., 1998, Galaxy Formation. Springer, New York

Misner C. W., Thorne K. S., Wheeler J. A., 2002, Gravitation. W. H. Freeman and Company, San Fransisco

Perlmutter S., et al., 1997, ApJ 483, 565

Perlmutter S., et al., 1999, ApJ 517, 565-586

Ruhl J. E., et al., 2003, ApJ 599, 785-805

Sahni V., 2004, arXiv:astro-ph/0403324

Schramm D. N., 1990, in Astrophysical Ages and Dating Methods, ed. E. Vangioni-Flam et al., 365

Turner M. S., 2002, Int. J. Mod. Phys. A17S1, 180-196

Weinberg S., 1972, Gravitation and Cosmology. Wiley, New York

Zwicky F., Rudnicki K., 1963, ApJ 137, 707

Dark Energy and Quintessence in the Universe

Daniel Johansson

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(f00dajo@dd.chalmers.se)

*

Abstract

Recent cosmological observations indicate that the expansion of our universe is accelerating. The only way to account for such an acceleration is to introduce a new type of energy called Dark energy. The dark energy has a negative equation of state, meaning that its pressure is negative. The current observations imply that 70% of the energy density of the universe is dark energy.

The simplest form of dark energy is the so-called cosmological constant. It was introduced by Albert Einstein in the 1910's. The cosmological constant model of dark energy fits the observations well, but there are some theoretical problems with this model. The theoretical value of the cosmological constant is 124 orders of magnitude larger than the observed value. There is also the coincidence problem, which is the fact that today the fractional energy densities in dark energy and dark matter are comparable. It seems that we are living in a very special time of the universe.

There are also dynamical models of dark energy. *Quintessence* is a scalar field that is rolling down its potential. Quintessence might solve the coincidence problem. The *Brane-World*-models suggest that we live in a four-dimensional brane of the bulk, which is $(4+d)$-dimensional. There are also models that provide a smooth transition between matter- and dark energy domination. One is the *Chapyglin gas*. These models provide a better theoretical understanding of dark energy, although observations are still not good enough to exclude neither the cosmological constant nor dynamical dark energy.

1 Introduction

Cosmology is the observational theory of the large scale structure of the universe. The interest in cosmology has increased during the past years, mainly because of new high precision measurements. It is today possible to constrain the values of the important cosmological parameters to very high accuracy. But this is not the end of cosmology. In 1998, data were published that showed that the universe is filled to about 70% with an

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

unknown energy component. This form of energy has been named Dark Energy. It is the subject of this paper.

The layout of the paper is as follows: first some important definitions will be made in section 1.1. In section 1.2 the first theoretical thoughts of dark energy will be presented. This section also gives an introduction to the standard model of cosmology. In section 1.3 the Hubble expansion of the universe is explained. Section 1.4 is an introduction to the cosmic microwave background radiation and the big bang. In section 1.5 the observations that showed that there is an unknown energy component are presented. Then in section 2 dark energy is introduced, and a number of models of dark energy are discussed. The paper ends with a discussion in section 3.

1.1 Some Definitions

The redshift, z, is related to the scale factor of the universe $a(t)$ through

$$1 + z = \frac{a(t_{obs})}{a(t_{em})}, \tag{1}$$

where t_{obs} and t_{em} are the times for the observation and emitting of light.

The Planck units are combinations of the three fundamental constants c, the speed of light in vacuum, $\hbar$, Planck's constant divided by 2π and G, Newton's constant of gravitation. The Planck density, (which will be needed in section 2) ρ_{pl} (in units of $(\mathrm{kg\,m^{-3}})$), has the form

$$\rho_{pl} = \frac{c^5}{\hbar G^2} \simeq 10^{96}\ \mathrm{kg\,m^{-3}}. \tag{2}$$

Hence the energy of the Planck scale is much larger than any energy scale tested in laboratories.

1.2 General Relativity and the Cosmological Constant

In 1917, Albert Einstein published his theory of General Relativity, (GR). It is a generalisation of his theory of special relativity, and describes gravity. Einstein's brilliant idea was to connect the matter- and energy density to the curvature of space-time. Matter and energy *curves* space-time in such a way that particles moving in space-time feel a gravitational force because of the curvature. This is clearly an astonishing fact! The *Einstein equation* is the foundation of the theory. Setting c, the speed of light, equal to one (as is customary in relativity), the equation reads (Bergström & Goobar, 1999)

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{3}$$

It is a tensor equation with two components; $G_{\mu\nu}$, the *Einstein tensor*, describes the curvature of space-time. $T_{\mu\nu}$, the *Stress-Energy tensor*, describes the energy and matter density of space-time. G is Newton's gravitational constant.

Einstein's theory was first taken with large scepticism by the physicists of that time. But in 1919 a solar eclipse occurred. One of the implications of the theory of General Relativity is that light rays are bent by massive objects. This effect was measured during the eclipse, and the measured value of the bending agreed with the value calculated using GR. The success and fame of Einstein was instantaneous. But there were still some problems with the theory, problems of a more philosophical kind.

Einstein thought that the universe was static, i.e. that it has and always will be the same. This was the general conception at the time, it was not based on any observational evidence. Einstein wanted his equations to give this answer when solving them for the whole universe. The space-time metric used in cosmology is based on two main assumptions; *Isotropy* and *Homogeneity*. These assumptions are impossible to check exactly, but observation tell us that at large scales, the universe is both homogenous and isotropic. This leads us to the space-time metric called the *Friedmann-Lemaitre-Robertson-Walker*-metric (FLRW):

$$ds^2 = dt^2 - a(t)^2\Big(\frac{dr^2}{1-kr^2} + r^2(d\theta^2 + \sin^2\theta d\phi^2)\Big), \tag{4}$$

where (t, r, θ, ϕ) are polar four-coordinates, and ds is the line element. $a(t)$ is the scale-factor of the universe at time t. k is a constant related to the overall geometry of the universe. If $k = +1$ the universe is said to be *closed*. For $k = 0$ the universe is *flat* and for $k = -1$ the universe is *open*. We should point out that this metric is not found by solving the Einstein equation, but relies *only* on the assumptions of isotropy and homogeneity.

If we solve the Einstein equations for this metric, the result is the Friedmann equations (Bergström & Goobar, 1999):

$$\Big(\frac{\dot{a}}{a}\Big)^2 + \frac{k}{a^2} = \frac{8\pi G}{3}\rho, \tag{5}$$

and

$$2\frac{\ddot{a}}{a} + \frac{k}{a^2} = -8\pi G\rho. \tag{6}$$

Here the dots mean time derivatives and ρ is the sum of the densities of matter (ρ_M) and radiation (ρ_{rad}).

The problem with these equations is that they are not static. To fix the problem with a non-static solution, Einstein started to investigate ways to modify his equation. He found that the only possible modification that was permitted, was to add a constant term proportional to the metric tensor $g_{\mu\nu}$ (since the Einstein tensor has to be covariantly constant). He introduced the *Cosmological Constant* Λ and added it to the LHS of eq. (3) so that it read

$$G_{\mu\nu} - \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{7}$$

This term also changes the Friedmann equations in such a way that $\Lambda/3$ is added to the right-hand-side of eq. (5), and Λ is added to the RHS of eq. (6). We can then absorb the constant into the density by defining $\rho_\Lambda = 3\Lambda/8\pi G$ and

$$\rho_{tot} \equiv \rho_M + \rho_{rad} + \rho_\Lambda.$$

Hence, we can think of Λ as a constant energy density.

An important parameter in cosmology is the critical density, ρ_{crit}. It is related to the Hubble constant H introduced in section 1.3 through (Bergström & Goobar, 1999)

$$\rho_{crit} \equiv \frac{3H^2}{8\pi G}. \tag{8}$$

The physical interpretation of H_0 is also discussed in section 1.3. From the critical density we define the density parameter $\Omega \equiv \rho/\rho_{crit}$. Ω is related to the curvature parameter k and the scale factor a through

$$1 - \frac{k}{2a^2} = \Omega. \tag{9}$$

The physical interpretation of the critical density is that if the density of our universe is equal to the critical density, the universe is spatially flat, hence $k = 0 \Leftrightarrow \Omega = 1$.

The *equation of state*, w, is the relation between pressure and density. It is defined as $w \equiv p/\rho$. We derive an important equation by taking the difference between eq. (6) and eq. (5) and using the equation of state. The result is called the acceleration equation

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) = -\frac{4\pi G}{3}\sum_i \Big[\rho_i(1 + 3w_i)\Big], \tag{10}$$

where the summation over i refers to different types of energy. From equation (10) we can derive an important relationship between ρ and a.

$$\rho \propto a^{-3(1+w)} \tag{11}$$

The three main types of energy in the universe are radiation, matter and the cosmological constant energy. Using eq. (11) we see that for matter, $w = 0$ (the matter in the universe has no pressure). For radiation (the Cosmic Microwave Background, CMB), $w = 1/3$. For the cosmological constant, ρ is constant and $w = -1$. Hence, the constant energy density of Λ gives rise to a *negative* pressure!

When Einstein saw that the adding of the cosmological constant made the solutions static, he was satisfied. But the picture was soon going to change.

1.3 Hubble's Discovery

In 1929, Edwin Hubble published data that showed that the universe was *not* static, but that it was expanding. He had observed galaxies and found that their rescession speed v was proportional to their distance to us (d). He formulated his discovery in the famous Hubble law

$$v = Hd, \qquad H \equiv \frac{\dot{a}}{a}, \tag{12}$$

where the proportionality constant H is called the Hubble constant. It has the dimension of time^{-1} and is related to the age of the universe. To accurately measure H has proven to be a hard task, and still today there is a large uncertainty in this for cosmology extremely important constant. The value is $H = 100 \cdot h\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$. Today the parameter h has the value $h = 0.72 \pm 0.8$ (Freedman, 2001). The expansion of the universe in this model is global, which means that everything is moving away from everything. It can be understood by the "balloon"-model: imagine that galaxies are sitting on the edge of a balloon as tiny dots. We blow air into the balloon so that it expands. Then all the dots will move away from each other. Our place in the universe is not special; if we would look from some other direction we would see the same pattern of expansion.

When Einstein saw Hubble's result he realized that his idea of a static universe was wrong and he called the introduction of Λ his "biggest blunder". But people did not forget about the cosmological constant; it has been reintroduced at different times for different purposes. The first reintroduction of the cosmological constant was a few years after the discovery of expansion. Hubble had found $h \simeq 5$ in his measurements, which indicated that the age of the universe would be smaller than the age of the earth. If the cosmological constant was reintroduced it could make the universe older and solve this

problem. But Hubble's first measurements of the recession speed of galaxies were not very accurate. His value of the Hubble constant was ten times too large. So the cosmological constant was abandoned again. Today, observations tell us that Λ actually is nonzero. This is the subject of section 2.

1.4 The Big Bang and the CMB

In the standard model of cosmology, the universe starts with the big bang; a huge explosion from a singularity. Hence, the universe starts to expand from a small point in space-time. Right after the big bang, the temperature was so high that elementary particles (maybe even quarks) were free. The expansion caused the universe to gradually cool. In the very first moments after the big bang, we have no physical theory to describe the universe. It seems that at these high energies, quantum gravity is needed. Still, there is no such theory, although the theories that are being developed at this moment (i.e. string theory and M-theory) seem promising.

The universe continued to cool, hence reducing the energy. During these first moments, different species of particles "freezed in" when the temperature became lower than their rest energy. At a redshift of $z \sim 1100$, the temperature was so low that the hydrogen ionization energy exceeded the mean energy of photons. Hence hydrogen became stable and the photons were free to travel through the universe. These photons were distributed according to a blackbody spectrum. When the universe continued to expand, the temperature of the photons decreased.

In 1963 Penzias and Wilson, two engineers at the Bell laboratories in the US, were testing a new antenna. Their measurements were disturbed by a microwave- frequency noise. After discussing this with cosmologists, they realized that they had found the Cosmic Microwave Background (CMB) radiation. Penzias and Wilson received the Noble Prize in physics for their discovery, although the explanation of the measurements was not their own. Today the spectrum of the CMB is measured to great precision. It is an almost perfect blackbody spectrum of temperature $T \simeq 2.7$K.

1.4.1 Inflation and Anisotropies in the CMB

The big bang model seems to describe the universe fairly well. But the standard big bang model has some problems that for some time troubled cosmologists. First, the CMB is to great precision *isotropic*, i.e. it looks the same in all directions. The problem is: how is it possible that different parts of the sky have the same temperature? If the universe started from a small, dense region and then expanded, there would not be enough time for the system to reach thermal equilibrium because of the expansion. Since the speed of light is finite, different parts of the universe would not have a chance to communicate with each other and reach thermal equilibrium.

In the 1970's, cosmologists started to investigate possible ways to make the CMB isotropic. The concept of *inflation* was proposed in the beginning of the 1980's, and is today a part of the standard model of cosmology. The main idea about inflation is that a few moments after the big bang, the universe went through a short phase of exponential expansion. For $k = 0$ and $\Lambda > 0$ the scale factor (which is calculated using the Friedmann equations) will have the time dependence $a(t) = \exp(\sqrt{\Lambda/3}t)$ (Bergström & Goobar, 1999). One of the main properties of inflation is that it will make $k = 0$. This can be understood if we consider a spherical surface that is expanded. Finally the surface will appear flat if it has been expanding for a long time.

The second problem with the CMB is that it must be *anisotropic* on small angular scales! How is this possible? If the temperature of the universe would be exactly the same everywhere right after the recombination, there would be no density fluctuations and no structures would be able to form. Measurement made by the COBE satellite in the 1990:s confirmed the predictions. The CMB is anisotropic on small angular scales.

The multipole index is l is used to characterize the anisotropies in the CMB. It is (approximately) inversely proportional to the angular scale $\Delta\theta$, i.e. $l \propto 1/\Delta\theta$. Hence, at large l the angular scale is very small. Measurements of the temperature fluctuations at $l > 100$ were made during the late 1990:s. The result of these measurement is shown in this graph, where the temperature fluctuations ΔT are plotted as a function of l.

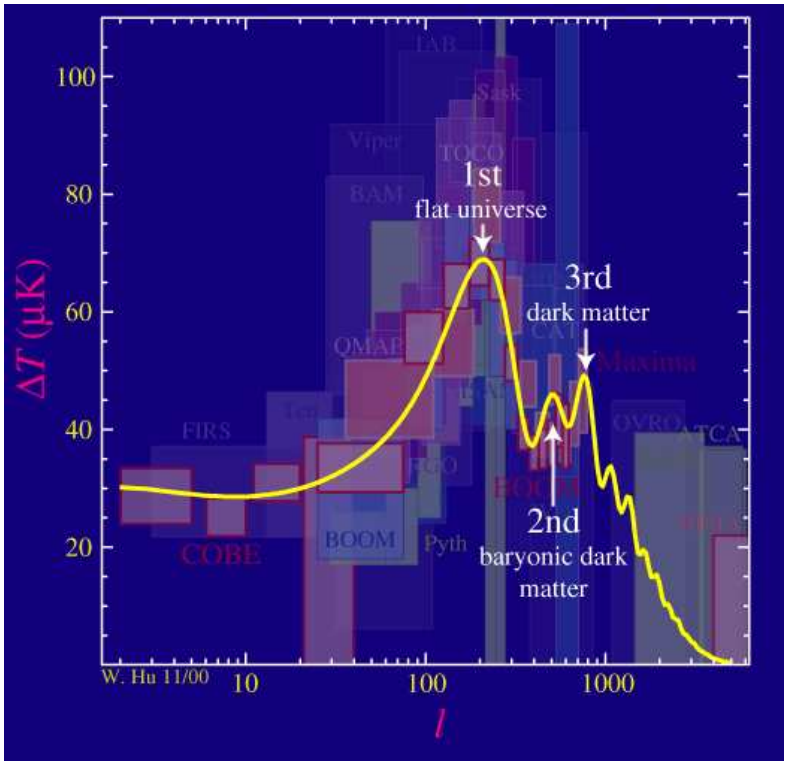

Figure 1: The Temperature-multipole spectrum. From Hu (2004).

Theoretical models of the temperature anisotropies indicated that the spectrum would peak at $l \approx 200$ (Hu, 2004). As we see in the graph above, $l \approx 220$. A truly astonishing feature of the power spectrum of the temperature anisotropies is that it can be used individually to accurately determine the values of cosmological parameters. For example, the first peak in the power spectrum is related to the curvature of the universe. The measurements tells us that the curvature parameter k (which was introduced in section 1.2) is close to zero (Hu, 2004). Hence the universe is nearly flat on enough large scales. $k = 0$ also corresponds to an spatially infinite universe. The second and third peak in the spectrum can be used to determine the number of dark baryons and the density of dark matter respectively (Hu, 2004).

1.5 The Expansion of the Uniiverse Is Accelerating

Hubble's model of the expanding universe has $\dot{a} \neq 0$, but $\ddot{a} = 0$. Hence, in his model the universe is expanding at a steady state, with no acceleration. In 1998, two individual research teams published results on supernova 1A measurements that showed that the

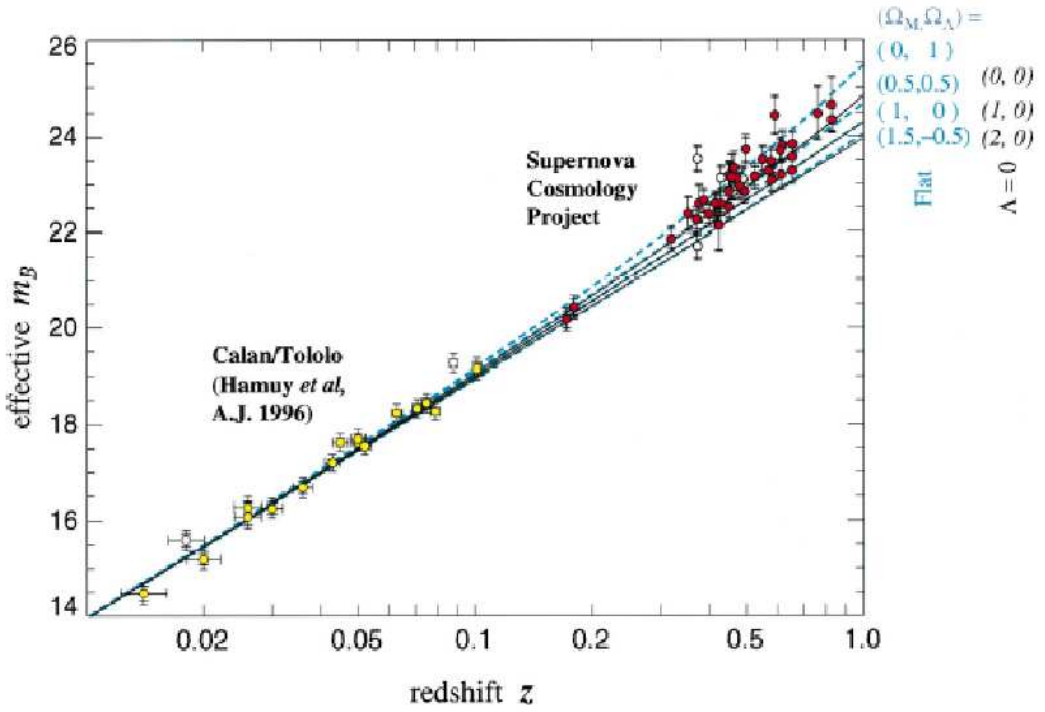

Figure 2: Hubble diagram of supernovae of high redhift. We see that the best fit to the data is *not* a straight line (as in the Hubble model), but a nonlinear relation. This indicates that the expansion of the universe is accelerating. (From Perlmutter et al. 1999).

expansion of the universe actually is accelerating. This can be understood from equation (10): if $1 + 3w < 0$, $w < -1/3$, and $\ddot{a} > 0$. The measurements are described in detail in Martin Nord's paper. Here I will only give a brief description on how the analysis of Supernova 1A data was used to determine that $\ddot{a} > 0$.

First I will discuss the problem of measuring distances in astronomy. For nearby astronomical objects we can use the trigonometric parallax to determine the distance. In our solar system we can even use radar methods to measure distances very accurately. For far away objects (e.g. galaxies at high redshift) the best method of measuring the distance is to use a *standard candle*. It is believed that some objects (e.g. SN1a) all have almost the same luminosity. This has been found by observing supernovae in nearby galaxies. Hence, they can be used to measure cosmological distances at high redshifts, because if they are less luminous, they must be farther away. A problem with this method of distance measurement is that the light observed at the earth has been obscured by interstellar and intergalactic gas and dust. But there is a way to resolve this problem. SN1a are exploding stars that apart from an almost identic luminosity also has a light curve (time plotted versus luminosity) that can be used to "normalize" the luminosity. Using this fact, astronomers can determine the distance to the supernova to an accuracy of about 20% , which is rather good with astronomical standards.

The observations in 1998 revealed that high redshift Supernovae 1a actually appeared *dimmer* than they should with the ordinary Hubble model. When plotting d versus v for a sample of galaxies, we expect to find a straight line with slope H in the hubble model. But one of the two research teams (Perlmutter et al., 1999) found the Hubble diagram shown in figure

The fact that distant supernovae appear dimmer can only be explained by an accel-

erating universe. If the expansion speed of the universe has increased since the emitting of light, the wavelength of the light will stretch and hence it will appear dimmer. This is maybe one of the most important discoveries in all of science during the last decade, and has started a lot of research on the subject. What is the force that drives the universe apart?

As was discussed above, eq. (10) yields that for an accelerating universe, $w < -1/3$. The two known types of energy that contributes to the curvature are radiation and matter. These have both $w \geq 0$. Hence, there must be some other type of matter or energy that has a negative w and can compensate the other two types. This energy-type has been given the name *Dark Energy*, and it is the subject of section 2.

2 Dark Energy and Quintessence

It was found in the previous section that the expansion of the universe is accelerating and that Dark Energy that has a negative equation of state is causing the acceleration. But what is the nature of dark energy? The short answer is: noone really knows!

The theory of inflation suggests that our universe is spatially flat, i.e. $k \simeq 0$. The measurements of the CMB-anisotropies have confirmed this prediction. Hence, $\Omega \simeq 1$ and the density of the universe is close to the critical density. What is the value of the critical density today? It can be calculated using the definition, eq. (8). If $h = 0.73$, $\rho_{crit} = 1.0 \cdot 10^{-28}$ kg m^{-3}.

If the universe is at the critical density, the different forms of energy must contribute so that $\Omega \simeq 1$. Observations show that the contribution of visible- and dark matter is $\Omega_M \simeq 0.3$. To add up to one, we must have the dark energy density $\Omega_{DE} \simeq 0.7$. There is also a negligle contribution to Ω from the Cosmic microwave background radiation (Krauss, 1999).

The acceleration equation can be rewritten using the *deceleration*-parameter q which is defined according to Sahni (2004)

$$q \equiv -\frac{\ddot{a}}{H^2 a} = \sum_i \frac{4\pi G \rho_i}{3H^2}(1 + 3w_i) = \frac{1 + 3w_X \Omega_X}{2}, \tag{13}$$

where we have assumed a flat universe and defined Ω_X as the density of dark energy compared to the critical density of the universe. We have also neglected the contribution of cosmic microwave background to the energy density. This radiation has $\Omega_{rad} \sim 10^{-5}$. Observations tell us that today, $\Omega_X \sim 0.7$ and $\Omega_M \sim 0.3$. Thus the radiation can be neclected.

If the dark energy has a constant equation of state, we can find the redshift z_a where the transition between deceleration and acceleration occured.

$$(1 + z_a)^{-3w} = -(1 + 3w)\frac{\Omega_X}{\Omega_M}. \tag{14}$$

We can also find the redshift z_{eq} at which the densities in dark matter and dark energy were equal;

$$(1 + z_{eq})^{3w} = \frac{\Omega_M}{\Omega_X}. \tag{15}$$

With $\Omega_X = 0.7$ and $\Omega_M = 0.3$ we find $z_a = 0.73$ and $z_{eq} = 0.37$ for $w = -1$. Hence the acceleration of the universe is a recent phenomen and it appears that we live in a special time of the universe where the densities of dark matter and dark energy are comparable.

This is called the coincidence problem. Why would we live in this special time? A possible solution to the coincidence problem is presented in section 2.1.2.

In section 2.1, I will describe vacuum energy and how it may act as a cosmological constant. In section 2.2, I will discuss dynamical forms of dark energy.

2.1 Dark Energy = Vacuum Energy?

The theory of quantum mechanics is maybe the most important discovery in physics during the 20:th century. One of the cornerstones of quantum theory is the Heisenberg uncertainty principle. It states that the uncertainty of two non-commuting observables (which in quantum theory become operators) is finite. The most famous form of the uncertainty principle is $\Delta x \cdot \Delta p > \hbar/2$. It states that the position x and the momentum p cannot be measured simultaneously to infinite precision. The precision of the measurement is limited by Planck's constant $\hbar$, which has a magnitude of $\sim 10^{-34}$Js. Another form of the uncertainty principle is $\Delta E \cdot \Delta t > \hbar$, where E is the energy and t is time. This expression can be rearranged to

$$\hbar/\Delta E < \Delta t. \tag{16}$$

The interpretation of this equation is the following: for a time interval smaller than Δt, a particle with energy ΔE can be created from pure vacuum. These particles are called virial particles. They cause, for example, alterations in the energy levels of atoms for short times. The De-Broigle wavelength of these particles spans the entire wavelength range. The *Casimir*-effect is a way to measure the effects of the virial particles. The experiment is simple: two metal plates are placed close together. Because of the finite spacing of the plates, virial particles of longer wavelength than this spacing cannot be created between the plates. Outside the plates, particles of all wavelengths are created. Hence, more particles are created outside the plates than inside, thus giving rise to a force that pushes the plates together.

How can the virial particles be related to the cosmological constant and dark energy? Since these particles actually have real energies, they should contribute to the stress-energy tensor in the Einstein equation. In the 1960's, it was discovered that vacuum energy must obey Lorenz invariance, and therefore be of the form $\langle T_{\mu\nu} \rangle = \Lambda g_{\mu\nu}$. Taking the tt component of this equation yields

$$\Lambda = 8\pi G \langle T_{tt} \rangle \propto \int_0^\infty \sqrt{k^2 + m^2} k^2 dk, \tag{17}$$

where k is the wavenumber and m is the mass of the particles. As we see, this integral diverges as k^4. We can make a cutoff at the Planck scale, which leads to an energy density $\rho_{pl} \simeq 10^{96}$ kg m^{-3} (calculated in seection 1.1). Comparing this to the critical density of our universe, we see that the value is $96 - (-28) = 124$ orders of magnitude too large. This is called the "cosmological constant problem". In the two following sections two different approaches to solving this problem are presented.

2.1.1 Supersymmetry

With the development of supersymmetry in the 1970's, it was believed that the cosmological constant problem would be solved. Supersymmetry is a quantum field theory which predicts that every elementary particle in nature has a supersymmetric partner;

a so called "superpartner". None of these superpartners have been observed in experiments, which indicates that they should have massed well above $\sim 100\,\mathrm{GeV}$, which is the currently largest energy scale investigated in accelerators. The reason why their masses are so large is then that the supersymmetry is broken at low energies. Supersymmetry predicts a balance between bosons and fermions in nature. Since these types of particles contribute with different sign to the vacuum energy, the chance was that a fine-tuning of this balance would produce the value of Λ that is observed today. There is one problem with this model though. The temperature of the universe today is very low, making the energy low. Hence the supersymmetry would be broken in our universe today, making Λ very large. If we instead consider inflation, which requires a large Λ, the supersymmetry would still exist at the high energy at the time of inflation. Hence, supersymmetry seems to create the opposite scenario than what is needed.

2.1.2 Antrophic Selection

Some cosmological models imply that the big bang of our universe is only one of many big bangs. The different big bangs occur at points far from each other in space-time. The antrophic argument goes as follows; if the vacuum energy density ρ_Λ varies between all these big bangs, the vacuum energy density that is observed by any astronomer in any universe must be such that ρ_Λ permits the existence of life. That is, if ρ_Λ is too large, large scale structures will not be able to form because of the rapid expansion of the universe.

Since galaxies must be able to form, we can use a spherical infall model (Peebles, 1967) to find an upper bound on the vacuum energy density

$$\rho_\Lambda < \frac{500\rho_M\delta_M^3}{729}, \tag{18}$$

where ρ_M is the mass density and δ_M is the fractional density pertubation. Both observables are taken at the time of recombination.

But it is not enough to live in a big bang where galaxies are able to form, and which has ρ_Λ according to eq. (18). More likely is that we will live in a big bang where intelligent life is possible. The probability of a civilization observing a vacuum energy density between ρ_Λ and $\rho_\Lambda + d\rho_\Lambda$ is (Weinberg, 2000)

$$d\mathcal{P}(\rho_\Lambda) = \mathcal{N}(\rho_\Lambda)\,\mathcal{P}_{\text{a priori}}\,d\rho_\Lambda, \tag{19}$$

where $\mathcal{N}(\rho_\Lambda)$ is the average number of scientific civilizations in big bangs with energy density ρ_Λ and $\mathcal{P}_{\text{a priori}}$ is the *a priori* probability that a particular big bang has a vacuum density between ρ_Λ and $d\rho_\Lambda$.

How can one calculate $\mathcal{N}$ and $\mathcal{P}_{\text{a priori}}$? Weinberg (2000) argues that $\mathcal{N}$ is related to the number of baryons that end up in galaxies. Since the vacuum energy density is much lower than the usual energy densities in particle physics, $\mathcal{N}$ is only non-zero in a small range of ρ_Λ. In this range, $\mathcal{P}_{\text{a priori}}$ will then be constant. Using a spherical infall model (Gunn & Gott 1971), Martel, Shapiro & Weinberg (1998) calculated $\mathcal{N}$. Using this $\mathcal{N}(\rho_\Lambda)$, the integrated probability of finding a universe with vacuum energy density smaller than or equal toρ_Λ is

$$\mathcal{P}(\leq \rho_\Lambda) \equiv \int_0^{\rho_\Lambda} d\mathcal{P} = 1 + (1+\beta)e^{-\beta} + \frac{1}{2\ln 2 - 1}\int_\beta^\infty F(x)\mathrm{d}x, \tag{20}$$

where

$$\beta \equiv \frac{1}{2\sigma^2}\Big(\frac{729\rho_\Lambda}{500\rho_M}\Big)^{2/3}, \tag{21}$$

and

$$F(x) \equiv e^{-x}\Big\{ -2\sqrt{\beta x} + \beta + 2x \ln\left(\sqrt{\beta/x} + 1\right)\Big\}. \tag{22}$$

σ is the rms fractional density pertubation at recombination. From eq. (20), (21) and (22) we can calculate the probability of living in a big bang with $\Omega_\Lambda = 0.7$. The result is between 5% and 12% and depends on how σ is estimated. Thus, according to this model, it is possible for intelligent life to exist in a big bang with our current value of ρ_Λ. Our value of ρ_Λ may appear a little low though.

It it important to point out that anthropic arguments are only valid if the underlying assumptions are valid. Of course it would be better to find a way to calculate the energy density from scratch, but since that has been found troublesome, the anthropic argument offers a way to determine the values of parameters anyway. Similar arguments have been made in other parts of astronomy; Tegmark and Rees (1998) argue that due to antrophic effects, the CMB fluctuation level is 10^{-5}. A similar argument has to be made to determine the size of the planetary orbits in the solar system. The orbits cannot be determined using Newton's dynamics. But for life to evolve on earth, water has to be liquid. This sets a limit on the size of the earth's orbit, since if we were closer to the sun, the water would evaporate and if we were further away the water would freeze to ice, and life would not be able to form.

2.2 Dynamical Dark Energy

In the previous section the dark energy was assumed to be vacuum energy of constant density. When the universe expands, the density in matter and dark matter will decrease and eventually the vacuum density will dominate. But the two problems with the vacuum energy; the cosmological constant problem and the coincidence problem have not yet been solved. The next logical step is then to consider an equation of state that is time dependent; $w = w(t)$. It is equivalent to a time-dependent Λ. In section 1.2 it was mentioned that a constant Λ is the most general expansion of Einstein's theory. Thus, making Λ time-dependent means that we are abandoning Einstein's theory of gravity! There are numerous models of dynamical dark energy. Here a few are discussed.

2.2.1 Quintessence

The name quintessence is adopted from the ancient Greeks, who named their fifth "element quintessence. Quintessence is a scalar field Q which has the Lagrangian density

$$\mathcal{L} = \frac{1}{2}\dot{Q}^2 - V(Q), \tag{23}$$

$$\rho \equiv T^t_t = \frac{1}{2}\dot{Q}^2 + V(Q), \quad p \equiv -T^\alpha_\alpha = \frac{1}{2}\dot{Q}^2 - V(Q), \tag{24}$$

The Lagrangian yields the equation of motion of the quintessence field (Steinhart, Wang, Zlatev 1999):

$$\ddot{Q} + 3H\dot{Q} + V' = 0 \tag{25}$$

where a prime denotes derivation with respect to Q. The important feature of the quintessence field is that it "rolls" down its potential. This means that at early times in the universe the quintessence density is large, hence making inflation possible. As the universe evolves, the quintessence density decreases to its present low value as the field

is rolling down the potential. The first models of quintessence still had the problem of fine-tuning the initial value of the quintessence density to extrapolate to the present value. These models then had the same problem as the cosmological constant; it is very hard to explain why nature fine-tunes the value of the cosmological constant to one part in 124 orders of magnitude. In 1999, Steinhart, Wang and Zlatev introduced the so-called "tracker field", which is a quintessence field with a potential that has special properties. The tracker field has an equation of motion that has attractor-like solutions. The solu-

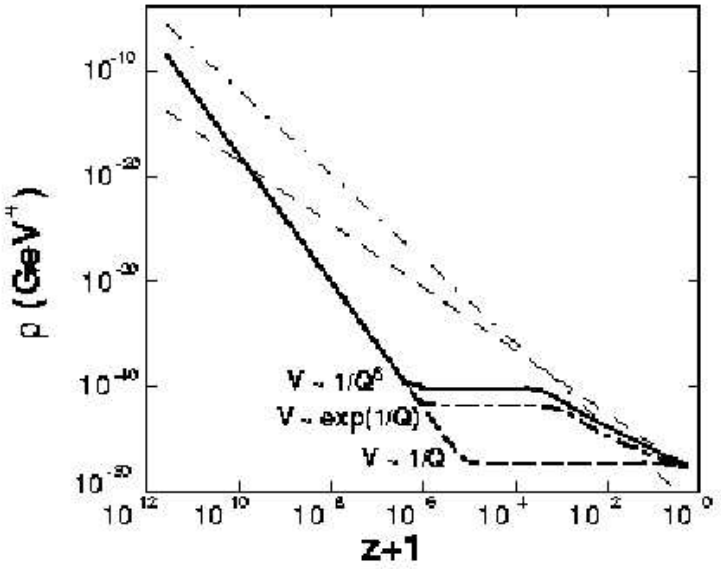

Figure 3: Tracker behaviour for the potentials $\sim 1/Q^6$ (top), $\sim \exp(1/Q)$ (middle) and $\sim 1/Q$ (bottom). The two upper potentials joins the tracker solution asymptotically. The lower potential does not tracker the tracker solution, because it violates the two conditions for tracking. From Steinhart, Wang and Zlatev (1999).

tions converge to a common path from different initial conditions. Steinhart, Wang and Zlatev (1999) find that the initial value of the quintessence density ρ_Q may vary by 100 orders of magnitude. This property of tracker solutions is desirable since inflation requires a large dark energy density, and today's energy density is very low.

The equation of state of the quintessence field is found from equation (24)

$$w_Q \equiv p/\rho = \frac{\frac{1}{2}\dot{Q}^2 + V(Q)}{\frac{1}{2}\dot{Q}^2 - V(Q)} \tag{26}$$

The equation of motion of the Q-field can, combined with the equation of state, be written

$$\pm\frac{V\prime}{V} = 3\sqrt{\frac{8\pi/3}{\Omega_Q}}\sqrt{1+w_Q}\Big[1+\frac{1}{6}\frac{d\ln x}{d\ln a}\Big] \tag{27}$$

where $x = (1+w_Q)/(1-w_Q)$. The tracking solution (to which solutions converge) has the property that the equation of state w_Q is nearly constant and lies between -1 and w_B, the equation of state of the background energy. The background energy differs from time to time; at early times the background energy is the radiation, which has $w = 1/3$. As the universe evolves, dark matter becomes the dominant energy form. As was discussed

in section 1.2, matter and dark matter has $w = 0$. Taylor expanding equation (27) to first order in w_Q yields

$$\frac{V'}{V} \approx \frac{1}{\sqrt{\Omega_Q}} \approx \frac{H}{\dot{Q}}, \tag{28}$$

where $\dot{Q} \approx \Omega_Q H^2$ is used in the last step. This is called the "tracker condition" for a tracker solution.

The function $\Gamma = \frac{V''V}{(V')^2}$ determines if tracker solutions exist for the potential $V(Q)$. For tracker solutions to exist, $\Gamma > 1$ and Γ has to be constant over the range of possible initial conditions. The last condition can be tested using the estimate

$$|\frac{\Gamma'}{\Gamma(V'/V)}| \ll 1. \tag{29}$$

To use this estimate, the expression is to be evaluated for the whole range of initial conditions of Q. If the inequality holds for all these values, a tracker solution exists, provided $\Gamma > 1$. These conditions apply when $w_Q < w_B$, which is the normal scenario. If instead $w_B < w_Q < (1/2)(1 + w_B)$, Γ must be smaller than one for a tracker solution to exist. Since Γ only depends on the potential V, it is not necessary to solve the equation of motion for the Q-field (equation (25)) to find out if a tracker solution exists for a certain potential.

Two simple potentials which have tracker behaviour are the inverse power potential $V(Q) \propto Q^{-\alpha}$, with $\alpha > 0$, and the exponential potential $V(Q) \propto \exp(M/Q)$, where M is a constant. Two potentials that *do not* exhibit tracking behaviour are $V(Q) \propto Q^2$ and $V(Q) \propto (\cos(Q) + 1)$. In these two models, a fine-tuning of the initial conditions is necessary. In figure 3, the tracker behaviour is showed for three different potentials.

2.2.2 Brane-World Models

In the brane-world models, the observable universe consists of a $1+3$ dimensional surface (the "brane") which is a part of a $1+3+d$-dimensional spacetime (the "bulk"). Particles of the standard model are confined to the brane, but gravity can propagate from the brane to the bulk. This is not surprising since gravity is the dynamics of space-time itself.

If the metric of the four dimensions is independent of the position in the additional dimensions, the size of the extra dimensions must be smaller than 1 mm to agree with observations. Otherwise, these dimensions would have resolved in tests of gravity. This is the case of *compact* dimensions. As opposed to this idea, Randall and Sundrum (1999) argue that there might be more than four noncompact dimensions. This happens when the metric of the brane is dependent of its position in the bulk. They show that a $4+1$-dimensional bulk space-time reduces to ordinary Einstein gravity. Also, they show that only gravitational radiation can disappear from the brane.

The equation of motion for a scalar field Q propagating on the brane is the same as eq. (25), with the new Hubble parameter (Sahni, 2004)

$$H^2 = \frac{8\pi}{3m^2}\rho(1 + \frac{\rho}{2\sigma}) + \Lambda_4/3 + \frac{\epsilon}{a^4} \tag{30}$$

where $\rho = 1/2\dot{Q}^2 + V(Q)$ and ϵ is a constant of integration related to gravitational radiation. The brane tension σ relates the four-dimensional Planck mass (m) to the five-dimensional Planck mass (M).

$$m = \sqrt{\frac{3}{4\pi}}(\frac{M^3}{\sqrt{\sigma}}). \tag{31}$$

σ also relates the four dimensional brane cosmological constant Λ_4 to the five-dimensional bulk cosmological constant Λ_b through

$$\Lambda_4 = \frac{4\pi}{M^3}\Big(\Lambda_b + \frac{4\pi}{3M^3}\sigma^2\Big). \tag{32}$$

There is also an extra term $\frac{\rho}{2\sigma}$ in eq. (30). This term occurs because of conditions at the boundary between the bulk and the brane. The term is important because a field rolling down the potential V will roll slower when the term is present. Hence, one can use steeper potentials and still get a slow rolling field. This fact has provided a probable link between inflation and quintessence since one can use the quintessence potentials also for inflation in this model.

2.2.3 Chapyglin Gas

The Chapyglin gas is a model of a gas with the non-standard equation of state (Kamenshchik, Moschella & Pasquier, 2001)

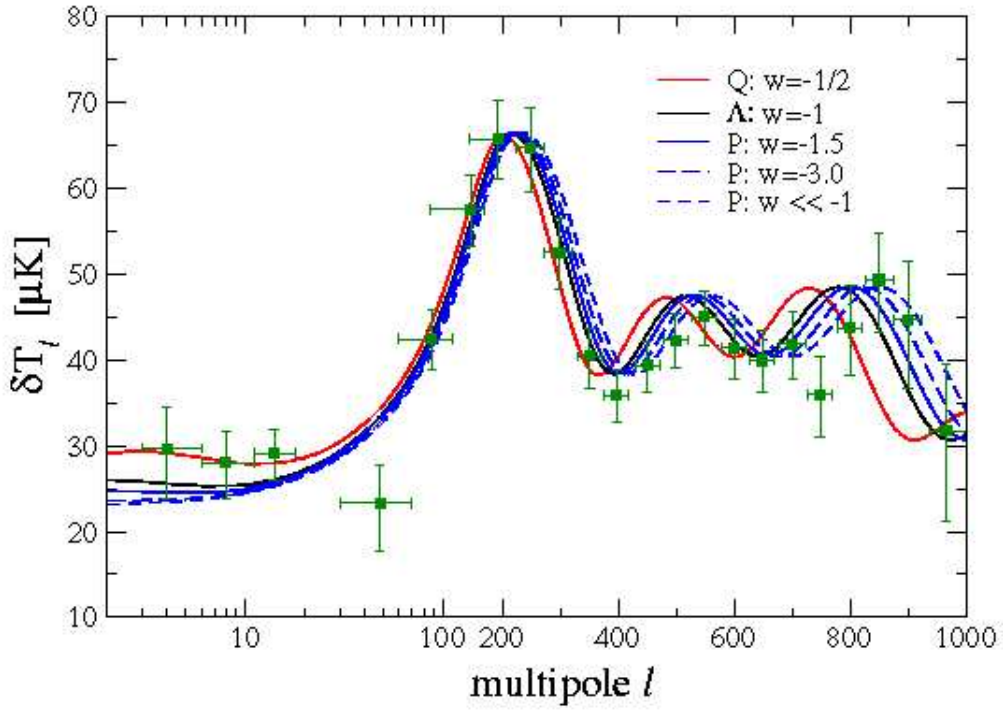

Figure 4: The CMB anisotropy spectrum for quintessence (first), cosmological constant (second) and phantom (three last). From Caldwell (2002)

$$p = -\frac{A}{\rho}, \tag{33}$$

where A is a positive constant. Energy conservation, $\mathrm{d}\,(\rho a^3) = -p\,\mathrm{d}\,(a^3)$, and the equation of state (33) then yields

$$\rho = \sqrt{A + \frac{B}{a^6}} = \sqrt{A + B(1+z)^6}. \tag{34}$$

Thus at early times the Chapyglin gas behaves like pressureless dust, and at later times it behaves like a cosmological constant. The Chapyglin gas therefore interpolates between

different eras in the evolution of the universe (Sahni, 2004). Another model that has this property is described in section 2.2.5.

2.2.4 Phantom Energy

The equation of state of the cosmological constant is $w = -1$. All other models discussed here have $w \geq -1$. Observations today are consistent with values down to $w = -1$. Caldwell (2002) considers a model with $w < -1$. He calls his model "Phantom Energy". From equation (11) we see that having $w < -1$ corresponds to an increasing energy density. Since the density of the phantom energy increases, the transition between matter domination and dark energy domination must have happened later than in other models. The scale factor of the phantom model is related to the scale factor at the end of matter domination t_m through

$$a(t) = a(t_m)\Big[(1+w)\frac{t}{t_m} - w\Big]^{2/3(1+w)} \quad \text{for} \quad t > t_m. \tag{35}$$

We see that this expression is divergent for $t = wt_m/(1+w)$. Hence, in this model, the universe diverges in a finite time. Caldwell then compares the phantom model to other models of dark energy. The CMB-anisotropy spectrum for the phantom model, quintessence and the cosmological constant. Also the redshift-magnitude dependence for SN1a samples is shown.

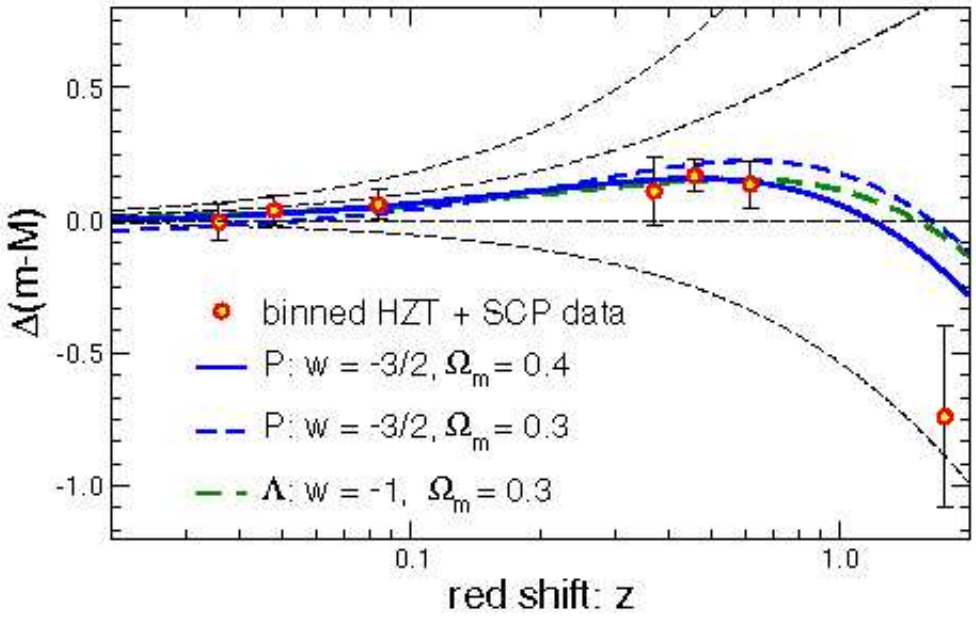

Figure 5: Magnitude-Redshift relation for two phantom models (see graph), and the cosmological constant. The dashed thin lines are not important here. Caldwell (2002).

Knop *et al.* (2003) find in a study of high redshift supernovae that if these measurements are combined with CMB measurements and 2dFGRS, the 95% confidence interval limits a constant equation of state to $-1.61 < w < -0.78$. Hence there is a chance that phantom energy actually exists.

Sahni, (2004), notes that the sound speed $v = \sqrt{|dp/d\rho|}$ can become larger than the speed of light in the phantom model.

2.2.5 Other Models

In 2004, Cardone, Troisi & Capozzeillo constructed a general class of models that have smooth transitions between radiation-, matter- and quintessence domination. These models are phenomenological, i.e. they are built on what we know of the universe and attempting to reproduce that. The usual scientific way would be to make a model that seems to be valid from the underlying physics and then try to reproduce observations. Cardone et al propose a model with a single fluid whose energy density scales with redshift in such a way that the different eras in cosmic evolution are extracted.

To mimic the cosmic evolution, the following expression for the energy density is assumed

$$\rho(R) = A\Big(1+\frac{s}{a}\Big)^{\beta-\alpha}\Big[1+\Big(\frac{b}{a}\Big)^{\alpha}\Big] \tag{36}$$

where $0 < \alpha < \beta$, and s, b two scaling factors with $s < b$. A is a normalization constant and a the scale factor of the universe. This expression can be rewritten using $a = (1+z)^{-1}$

$$\rho_z = A\Big(1+\frac{1+z}{1+z_s}\Big)^{\beta-\alpha}\Big[1+\Big(\frac{1+z}{1+z_b}\Big)^{\alpha}\Big] \tag{37}$$

with the definitions $z_s \equiv 1/s - 1$ and $z_b \equiv 1/b - 1$. From equation (36) we see that

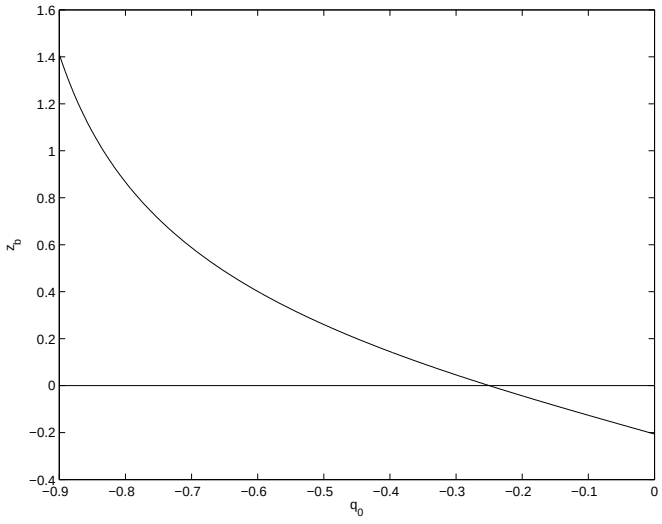

Figure 6: The dependence of q_0 with z_b for $(\alpha, \beta) = (3, 4)$ and $z_s = 3454$.

$$\rho \sim a^{-\beta} \quad \text{for} \quad a \ll s.$$

$$\rho \sim a^{-\alpha} \quad \text{for} \quad s \ll a \ll b.$$

$$\rho \sim const. \quad \text{for} \quad a \gg b.$$

Hence, for $\alpha = 3$ and $\beta = 4$ we have radiation dominance for $z >> z_s$, matter dominance for $z_b \ll z \ll z_s$ and a constant energy density is approached for $z \ll z_b$. Using the friedmann equations (eqs. (5) and (6)) and the contuinity equation $\dot{\rho} + 3H(\rho + p) = 0$, the redshift dependent equation of state can be calculated. The exact expression of $w(z)$ is not important, but its values in the different limits are interesting:

$$w \sim 1/3 \quad \text{for} \quad a \ll s.$$

$$w \sim 0 \quad \text{for} \quad s \ll a \ll b.$$

$$w \sim -1 \quad \text{for} \quad a \gg b.$$

which is exactly the behaviour of w that is needed for the right domination epochs.

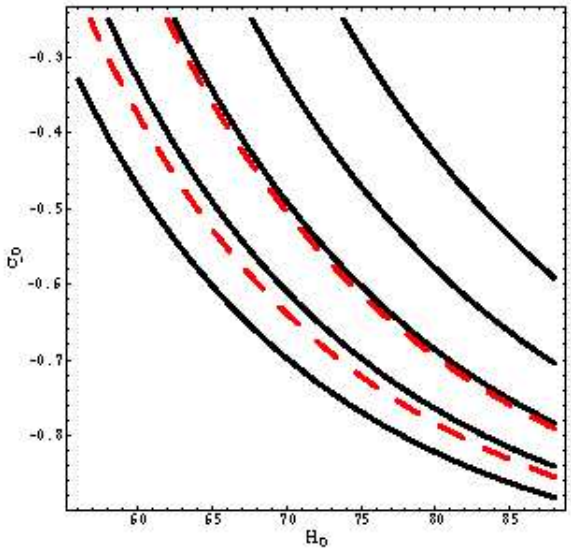

Figure 7: Contour plot of the age of the universe for H_0 and q_0. The dashed lines mark the region wich is supported by other observations. From Cardone, Troisi & Capozzeillo, 2004

Another relation that can be used to test the relevance of this model is the desceleration parameter $q = 1/2 + 3p/2\rho$ (here for a flat case) as a function of z_b. The present day value q_0 of the desceleration parameter has the dependence with z_b that is shown in figure 6.

The reason for choosing $z_s = 3454$ is that at this redshift the densities in radiation and matter are equal (Spergel et al. 2003). The authors also show that q_0 must lie within the range $[-1, -0.25]$, since the deceleration parameter is almost independent of z_b.

The model consists of five parameters, α, β, z_s, q_0 and H_0. The first three have already been fixed to 3, 4 and 3454. The last two can be constrained by using different cosmological data. Cardone, Troisi & Capozzeillo use three different test; the age of the universe, the hubble diagram of distant supernovae 1a and the angular size redshift test of high redshift radio objects. Contour plots of q_0 versus H_0 for the two first test are shown in figures 7 and 8.

The age of the universe test can not constrain H_0, but the constraint $q_0 > -0.85$ is found. For the Hubble diagram of SN1a, we see in the graph that it can't constrain q_0 but the following estimates can be made for the Hubble parameter (with the confidence levels written in parenthesis):

$$H_0 \in (58.8,\ 72.3)\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \quad (68\%\,\mathrm{CL})$$

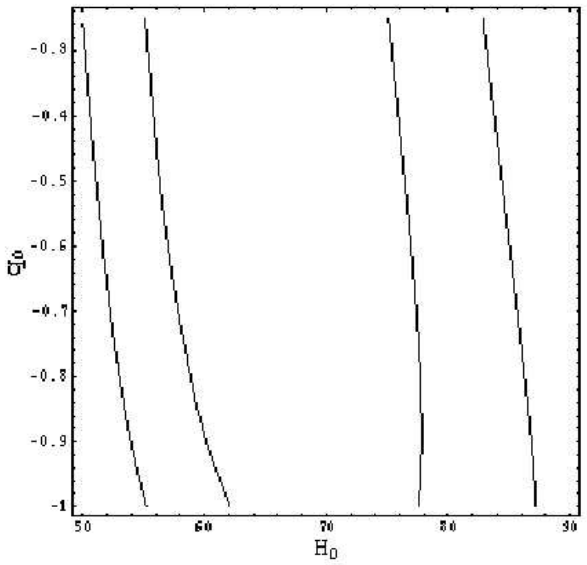

Figure 8: The 68% and 95% confidence intevals for fitting the model to the SN1a diagram. From Cardone, Troisi & Capozzeillo, (2004).

$$H_0 \in (53.1,\ 80.1)\,\mathrm{km\,s^{-1}\,Mpc^{-1}} \quad (95\%\,\mathrm{CL})$$

Both these ranges are in agreement with current observations. The angular size - redshift test sets the following constrains on q_0

$$q_0 \in (-0.76\ -0.54) \quad (68\%\,\mathrm{CL})$$

$$q_0 \in (-0.83,\ -0.37) \quad (95\%\,\mathrm{CL})$$

also in agreement with observations.

3 Discussion

The first section of this paper was devoted to understanding why the universe consists to a large extent of dark energy. In the second section, models of dark energy were considered. What can we say about these models? Can any of them be excluded by current observations?

The easy answer to this question is "no". Since the uncertanities in the present data are so high, all models fall inside the confidence levels. We see that some models fit some observations better than others, while the inverse can be true for some other type of observation. The only conclusion one can draw from this is that we need more observations! If the observations of high redshift supernovae continue to reduce the confidence intervals in determining $\Omega_{\mathrm{darkmatter}}$ and $\Omega_{\mathrm{darkenergy}}$, the models for dark energy can be further tested and improved. Also, more precise measurements of the power spectrum of the CMB will provide evidence if the universe really is flat or not. We see in figure 4 that even the unorthodox phantom energy fits the power spectrum better than the cosmological constant at high multipoles, although the opposite is true for the low multipoles. The magnitude-redshift relation of supernovae shown in figure 5 shows that the phantom energy actually fits the data very well if one allows the dark energy component to be $\Omega_{\mathrm{darkmatter}} = 0.4$. In figures 7 and 8 we see that the unification model of Cardone, Troisi & Capozzeillo, although phenomenological, can be used to constrain the cosmological parameters.

Of course there is a probabilty that the dimming of the supernovae is due to some other effect than acceleration, e.g. evolution of supernovae since the epoch of explosion, or maybe exctinction by the interstellar medium. But rescent observations have shown that even if these cases are taken into account, it is very probable that the acceleration of the universe is due to dark energy. But until it is proved that this is the case, one should always keep the other possibilities in mind.

To summarize, the following conclusions can be drawn:

- The universe is almost flat, and thus the density of the universe is close to the critical density.
- The universe consist to 30% of matter and dark matter, and to 70% of dark energy
- Dark energy has an negative equation of state and hence acts as a kind of anti-gravity
- Vacuum energy fits the observations well, but the theoretical value of the energy density is 124 orders of magnitude too high
- We also consider the anthropic selections, where we ask ourselves: "What value of the vacuum energy density permits life to form?" The answer that comes out of this analysis is remarkable since it is close to the observed value today. But instead of this anthropic method we would like to have a physical theory that explains dark energy.
- Dynamical dark energy: the equation of state is negative and time-dependent. Many models exist. Among the most popular is quintessence with tracker potentials, but many other are considered in the literature. These models also appear to be consistent with the observational data we have today.
- The future includes two different main objectives: firstly, higher precision observations are needed to constrain the values of the cosmological parameters. Secondly: theoretical work that explains dark energy and that fits to observations. Since the models tend to become more and more advanced, this field will likely be dominated by particle physicist in the future.

Acknowledgements

I would like to thank Martin Nord who helped me out a lot when discussing cosmology in general. I also would like to thank Cathy Horellou, who helped me getting started by providing some introductory articles and papers.

References

Bergström L., Goobar A., 1999, Cosmology and Particle Astrophysics. John Wiley & Sons Ltd, Chichester

Caldwell R.R., 2002, Phys. Lett. B, 545, 23, arXiv:astro-ph/9908168

Cardone V.F, Troisi A, Capozziello S, 2004 arXiv:astro-ph/0402228

Freedman W.L. 2001, Astrophys. J., 553, 47

Gunn, Gott J., 1972, Astrophys. J, 176, 1

Hu W. 2004 `http://background.uchicago.edu/~whu/intermediate/intermediate.html`

Kamenshchik A., Moschella U., Pesquier V., 2001, Phys. Lett. B, 511, 265

Knop *et al.* 2003, to be published in Astrophys. J.
arXiv:astro-ph/0309368

Krauss L.M., 1999, Scientific American, 1

Martel H, Shapiro P, Weinberg S, 1998, Astrophys. J, 492, 29

Peebles P.J.E., 1967, Astrophys. J., 147, 859

Perlmutter S. *et al.*, 1999, Astrophys. J., 517, 565

Randall L, Sundrum. R. 1999, Phys. Rev. Letters 83, 4690

Ratra B, Peebles P.J.E 1988, Phys. Rev. D 37, 3406

Sahni V., 2004, Lecture at the Second Aegean Summer School on the Early Universe, Syros, Greece, September 2003 (arXiv:astro-ph/0403324)

Spergel *et al.*, 2003, ApJS, 148, 175

Steinhardt P.J., Wang L, Zlatev I., 1999, Physical Review D, 59

Tegmark M, Rees M.J., 1998, Astrophys. J., 499, 526
arXiv:astro-ph/9709058

Weinberg S., 2000, Talk Given at Dark Matter 2000, Marina del Rey, CA, February 2000
arXiv:astro-ph/0005265

Starbursts in Merging Galaxies

Raquel Rodriguez Monje

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(raquel@oso.chalmers.se)

*

Abstract

The characteristics of normal and starburst galaxies are described, and the different triggers that lead to starburst activity are discussed. Such processes may be associated with interacting or merging galaxies but also with bars. Also, we describe active galactic nuclei (AGN) and the connection with starburst galaxies.

1 Introduction

The starburst galaxy together with its connection with the active galactic nuclei are two of the hot topics nowadays. In this paper, we introduce the different types of galaxies that exist, according to Hubble classification and how the interaction between then can lead to burst of star formation. There are other sources like bars that can trigger starburst.

The issue of a starburst-AGN connection in local and distant galaxies is relevant for understanding galaxy formation and evolution, the star formation and metal enrichment history of the universe, the origin of the extragalactic background at low and high energies, and the origin of nuclear activity in galaxies. In this paper we review some theoretical and observational evidence of a connection between the starburst and AGN phenomena.

2 Galaxies

2.1 Types of Galaxies

In his studies Edwin Hubble (1924) realized immediately that not all spiral galaxies have the same appearance. Furthermore, he found galaxies that do not have a spiral structure. Hubble classified the galaxies he studied according to their basic appearance. It was originally thought that the different types of galaxies represent different stages of galactic evolution. We now know that this is not the case. However, Hubble's classification scheme, depicted in Figure 1, is still quite useful.

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

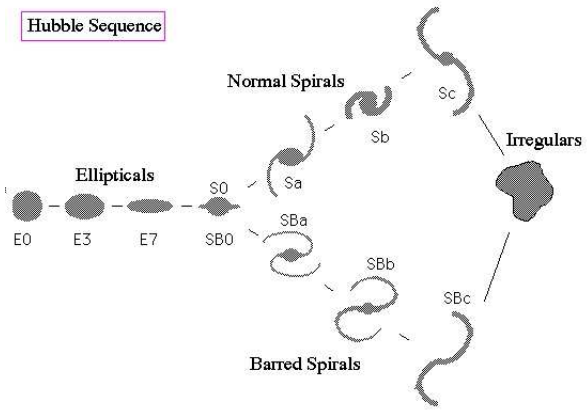

Figure 1: Hubble's Classification scheme.

Elliptical Galaxies: as their name suggest, have a simple elliptical appearance. The ellipticals are classified according to their degree of eccentricity. The ones that look spherical (zero eccentricity) are called E0, and the most eccentric are called E7. The most common types of elliptical galaxies are called dwarf ellipticals, since they are also the smallest. Their sizes are typically a few kiloparsecs and their masses are only a few million solar masses. More spectaculars are the giant ellipticals, with extents up to 100 kpc and masses of about 10^{12}Msun, with some masses up to a factor 10 higher.

Spiral Galaxies: spirals make up about two-thirds of all bright galaxies. They are subdivided into classes Sa, Sb and Sc. The two important features in the classification are: (1) the openness or tightness of the wrapping of the spiral pattern; and (2) the relative importance of the central bulge and the disk of the galaxy. Sa galaxies have the largest bulges and the tightest wound arms. Sc have the smallest bulges and the most open arms. Different types of spirals are shown in Figure 2. An important feature of spirals is the obvious presence of an interstellar medium – gas and dust. Even when a spiral is seen edge-on, we can tell if is a spiral by the presence of a lane of obscuring dust in the disk of the galaxy. The light from spirals contains an important contribution from a relatively small number of young blue stars, suggesting that star formation is still taking place in spirals.

Figure 2: Spirals galaxies. From left to right: M81 (Type Sa), M51 (Type Sb) and and NGC 2997 (Type Sc).

Some spirals galaxies have a bright bar running through their center, out to a point

where the arms appear to start. These are called *barred spirals galaxies*. They are similar to spiral galaxies in characteristics and in the classification of its sub-classes, SBa, SBb and SBc (see Figure 3). But a barred spiral possesses a bar of stars, as we have said before, typically of length a few kpc, running across the nucleus. The spiral arms originate from the ends of the bar rather that from the nuclear bulge. The Milky Way Galaxy was classified as a Sb or Sbc spiral for a long time. However, recent observations strongly suggest the existence of a bar. So its classification could change to SBb or SBc.

Figure 3: Barred Spirals galaxies. From left to right: NGC 3992 (Type SBa), NGC 1433 (Type SBb) and NGC 1300 (Type SBc).

There is an additional type of galaxy that is an intermediate type between spirals and ellipticals. They have a nuclear bulge disk but not spiral arms. This type is called *lenticular (S0) galaxies*. The bulge in an S0 is almost as large as the rest of the disk, giving the galaxy an almost spherical appearance. Some S0's also contain gas and dust, suggesting that they belong in the spiral classification. However, most S0's have no detectable gas. The role of S0's is still not well-understood.

Some galaxies have no regular pattern in their appearance. These are called irregular galaxies. They are typically small but rich in gas and dust. They may contain both young and old stars. We distinguish between two types of irregulars: Irr I are irregulars that appear slightly organized with a barely discernable disk or spiral arms. The Magellanic clouds are irregular companions of this type, to our own galaxy. Irr II galaxies just have a general amorphous appearance.

2.2 Interacting Galaxies

In the last two decades the concepts of galaxies as "island universes" that are slowly evolving virtually in total isolation have changed considerably. Since galaxies in a cluster are constantly in motion (because of the mutual gravitational force between them and their neighbors), from time to time they pass so close to one another that dramatic interactions can occur. These interactions come in a number of shapes and forms which depend upon a number of constraints. Some of these constraints are the difference of masses of the interacting, for instance, if the two galaxies are of similar masses then the results of the interaction are very different from when one galaxy is much larger than the other. Another possible constraints is the proximity of the galaxies, some galaxies pass one another and simply make their presence felt from a distance while others plunge together and merge into one.

If two galaxies pass each other at close range, then the tidal forces between the two galaxies will produce interesting distortions. For instance:

Bridges are formed when the target is much more massive than its companion. After a relatively direct passage of a small companion, the outer portions of the primary disk deform into a near-side spiral arm or "bridge", which extends toward the satellite, and a

far-side counterarm. A most famous example of a bridge is seen in M51 (the Whirlpool galaxy) and its companion, NGC 5195 (see Figure 4).

Figure 4: Left: ring Galaxy. Right: bridge (The Whirlpool galaxy).

When the masses of the two galaxies are equal or of the same order, the two internal spiral arms join up to form a relatively short-lived bridge while the two external arms are drawn out into long, curving *tails* of debris, which remain for one to two billion years. Figure 5 shows The Mice (Arp 242)

Simulations show that the tails, whose particles reach escape velocity, disperse into intergalactic space soon afterwards. Once detached from the mother galaxy, these dense complexes can form compact dwarf galaxies. Figure 4 shows The Antennae (NGC 4038/9), a well known pair of galaxies with tails. Interestingly, dense complexes of neutral hydrogen accumulate at the ends of the tails. These complexes also contain large quantities of HII gas, suggesting an increase in star formation.

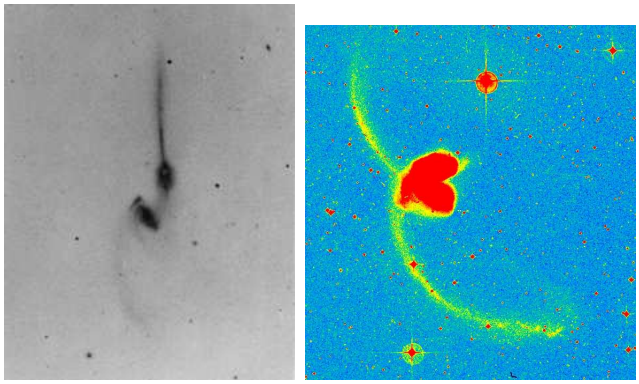

Figure 5: The Mice and the Antennae.

In the case of a close pass of a small compact galaxy with a large spiral, the large spiral will be relatively unaffected whilst the small compact galaxy will be radically altered. If, however, the compact galaxy actually passes through the spiral then it will cause the spiral to take on the shape of a *ring*.

When galaxies collide, the individual stars barely notice. The stars themselves never collide with each other, even in a galactic merger, because of the enormous distances between them, compared to their size. The effect of galactic interactions on the gas clouds, contained in galaxies, is rather different from the way the constituent stars are affected. Very often the new gravitational forces acting on the clouds trigger collapses that lead to an extremely vigourous burst of star formation. This phenomenon is known as a starburst.

2.3 Mergers

In recent years, a great deal of focus has been put on understanding merger events in the evolution of galaxies. Rapid technological progress in computers have allowed much better simulations of galaxies, and improved observational technologies have provided much more data about distant galaxies undergoing merger events. After the discovery in the last decade that our own Milky Way has a satellite galaxy (the Sagittarius Dwarf Elliptical Galaxy, or SagDEG) which is currently gradually being ripped up and "eaten" by the Milky Way, it is thought these kinds of events may be quite common in the evolution of large galaxies. The Magellanic Clouds are satellite galaxies of the Milky Way that will almost certainly share the same fate as the SagDEG.

When galaxies collide, the individual stars barely notice, as we have mentioned above. So when galaxies collide, they actually simply pass through each other, but the gravitational effects disrupts their structure as this happens. As they separate, gravity slows them down and, if they are gravitationally bound, will eventually bring them back together for another collision. After several collisions their individual structures are so changed, with many stars mixed up between them, that we identify the result as a single merged object. So after a merger, most of the stars originally belonging to both galaxies remain to form the new merged galaxy (a small fraction will have been thrown out entirely). If either galaxy were a spiral before the merger, very ordered system, the violence of event would disrupt the delicate structure of the disk. The existing stars cannot afterwards change their orbits to form a new disk. The stellar disk must essentially form in place; a dense rotating disk of gas forms first, then stars are born inside it.

3 Starburst Galaxies

Starburst are galaxies undergoing vigorous burst of star formation in their nuclei; the star formation takes place at a rate far higher that the average during a galactic lifetime. Starburst nuclei are detected optically by the presence of strong emission lines and blue continuum colors produced by hot stars. They also exhibit excessive infrared radiation attributed to dust heated by hot stars and radio emission from supernova remnants arising from massive stars. There is mounting evidence from simulations that burst of star formation may be induced by mergers of galaxies but gas flows in barred galaxies may also trigger the burst. Since stars form from interstellar matter, and from molecular clouds in particular it is important to make a short introduction to how the stars are form and what are the molecular clouds.

3.1 Molecular Clouds

Molecular clouds constitute the densest parts of the interstellar medium in galaxies, and ever since their discovery 25 years ago they have been of prime interest as the sites of star formation. One fundamental characteristic of molecular clouds is that they are not, as has sometimes been assumed, isolated "billiard balls" moving about independently in space, but instead are just dense condensations in more widely distributed, mostly atomic gas. Although molecular clouds may often appear to have sharp boundaries, these boundaries do not represent the edge of the matter distribution but just rapid transitions from the molecular gas to the surrounding atomic gas, which is distributed in extended envelopes that typically have comparable mass. Another important characteristic of molecular clouds is that they are transient structures and do not survive without

major changes for more that a few times 10^7 years.

The short lifetimes of molecular clouds are directly indicated by the fact that the range in age of the young stars associated with them is only about 10 to 20 Myr, comparable to the internal dynamical timescales of large molecular clouds; A third notable property of molecular clouds is that they are highly irregular structures ad have complex shapes that do not at all resemble equilibrium configurations.

Studying the molecular clouds in galaxies provides useful information on galaxy evolution and structure: (a) Molecular clouds collapse to form stars and their presence is an important signature of the starforming potential of a galaxy.(b) Since the interstellar matter (ISM) also serves as a waste-dump for processed material, its elemental abundances will reveal the nature of the past stellar generations.(c) Molecular clouds is a dissipative component in galaxy dynamics, it can thus significantly speed the merging process of some galaxy encounters.

The main components of the molecular clouds is Hydrogen. Although they also contain other molecular and atomic species, e.g. He, HI, CO. Because molecular hydrogen has no permanent electric dipole moment and the lowest quadrapole rotational transition lie in the infrared, most molecular line observations of galaxies are conducted in the J=1-0 rotational line of CO, the next most abundant molecular species. The first galaxies in which CO was detected were generally active galaxies. Although with improved receiver sensitivity many more normal galaxies are detected, the starburst galaxies remain among the galaxies with the highest CO luminosities.

3.2 Star Formation

Stars form inside molecular clouds. These regions are extremely cold (temperature about 10 to 20K, just above absolute zero). At these temperatures, gases become molecular, the group together. CO and H2 are the most common molecules in interstellar gas clouds. The deep cold also causes the gas to clump to high densities. When the density reaches a certain point, stars form.

Since the regions are dense, they are opaque to visible light and are known as dark nebula. Since they don't shine by optical light, we must use IR and radio telescopes to investigate them.

Star formation begins when the denser parts of the molecular cloud core collapse under their own weight/gravity. These cores typically have masses around 10^4 solar masses in the form of gas and dust. The cores are denser than the outer cloud, so they collapse first. As the cores collapse they fragment into clumps around 0.1 parsecs in size and 10 to 50 solar masses in mass. These clumps then form into protostars and the whole process takes about 10 millions years.

Once a clump has broken free from the other parts of the cloud core, it has its own unique gravity and identity and we call it a *protostar*. As the protostar forms, loose gas falls into its center. The infalling gas releases kinetic energy in the form of heat and the temperature and pressure in the center of the protostar goes up. As its temperature approaches thousands of degrees, it becomes a IR source.

The protostar, at first, only has about 1 percent of its final mass. But the envelope of the star continues to grow as infalling material is accreted. After a few million years, thermonuclear fusion begins in its core, then a strong stellar wind is produced which stops the infall of new mass. The protostar is now considered a young star since its mass is fixed, and its future evolution is now set.

This early phase in the life of a star is called the T-Tauri phase, where the star has:

vigorous surface activity (flares, eruptions)
strong stellar winds
variable and irregular light curves

A star in the T-Tauri phase can lose up to 50 percent of its mass before settling down as a main sequence star, thus we call them pre-main sequence stars. Their location on the HR diagram is shown below (Figure 6):

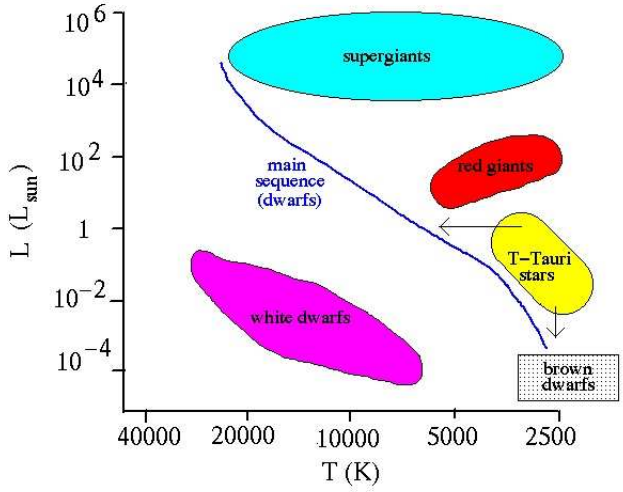

Figure 6: The HR diagram.

The arrows indicate how the T-Tauri stars will evolve onto the main sequence. They begin their lives as slightly cool stars, then heat up and become bluer and slightly fainter, depending on their initial mass. Very massive young stars are born so rapidly that they just appear on the main sequence with such a short T-Tauri phase that they are never observed.

3.3 Some Starburst Galaxies

3.3.1 M82 and NGC 253: Prototypical Starburst Galaxies

The two IR-bright galaxies M82 and NGC 253 are both classical starburst galaxies with strong radio continuum emission. Their proximity (3 Mpc) makes them ideal for detailed studies of the starburst phenomenon and its mechanisms since the starburst ISM can be studied at high linear resolution. Figure 7 shows a X-Ray picture of the M82 and an optical and X-Ray image of the NGC 253.

Although M82 is part of the M81, M82, NGC 3077 group of galaxies which shows evidence of interactions with H I tidal tails extending well beyond the optical discs of the galaxies and NGC 253 does not appear to have a companion, their nuclear sources are very similar. Observations of the radio continuum of M82 showed that the source breaks up into more than 30 compact sources, assumed to be young supernova remnants (SNR), immersed in a weak diffuse background, and that their combined luminosity was sufficient to explain the radio emission. There is no apparent evidence for a central AGN.

Observations of the CO (1-0) emission in M82 showed that most of the CO emitting gas could be located in molecular spiral arms 125 pc and 390 pc from the nucleus. The CO associated with the outer arm shows considerable velocity dispersion and has been disrupted, probably by starburst. The fact that the outer arm is disrupted is taken to

indicate that the older starburst is located in this arm and that the younger starburst took place in the inner arm recently enough that the surrounding interstellar medium has not been severely disrupted. Other authors like Shen and Lo (1995) have argued that if this is the case, the starburst is propagating inwards towards the center of the galaxy.

In contrast to M82, the molecular gas in NGC 253 shows a central molecular bar. Observations by Peng et al. (1996) show that the dense gas lies within a radius of about 300 pc. They speculate that the bar gets to channel gas into the central region of the galaxy, fuelling the starburst activity.

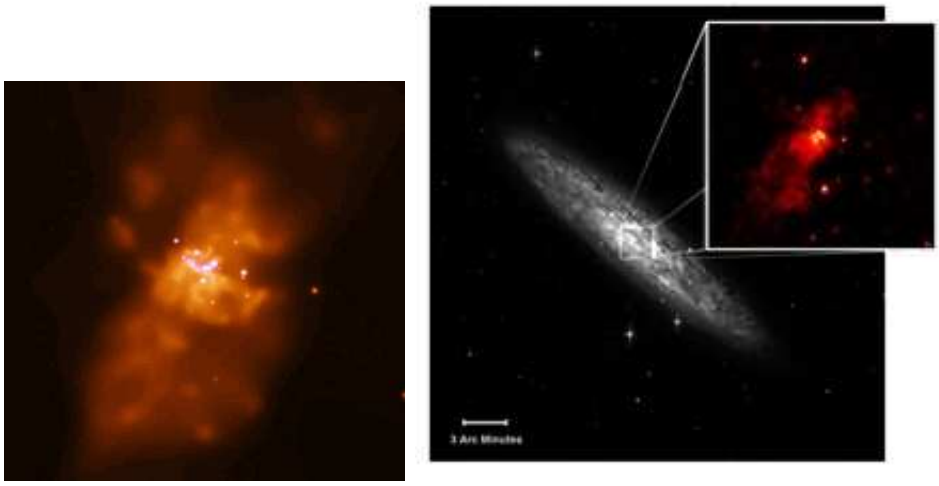

Figure 7: The Starburst Galaxies M82 (left) and NGC 253 (right).

3.3.2 NGC 3256: Starbust in Merging Galaxies

NGC 3256 is a starburst galaxy in which two extended tidal tails revel that an interaction between galaxies has occurred (see Figure 8). Kinematic information about the gas in such a system could help answer questions such as whether merger triggers starburst, whether starburst play a powerful role in active galactic nuclei, and whether mergers create gas concentrations.

NGC 3256 consists of two galaxies which are currently merging. Its starburst nature is indicated by data over a wide wavelength range:

The UV spectrum (Kinney et al. (1993)) has strong absorption features which indicate the presence of hot young stars;

The [Fe II] and radio continuum data (Norris and Forbes (1995)) indicate that the radio emission is due to synchrotron emission from cosmic-ray electrons accelerated by supernovae whose progenitors were massive stars.

Infrared observations by Graham et al. (1984), Doyon, Joseph, and Wright (1994), and Moor- wood and Oliva (1994) indicate that the K band continuum emission and CO strength are due to red supergiants and that the high Br line emission is due to OB stars ionising H II regions.

Smith and Kassim (1993) show that the spectrum of NGC 3256 closely resembles the archetypal starburst galaxy M 82 over the wavelength range from about 30 cm to 7500 A.

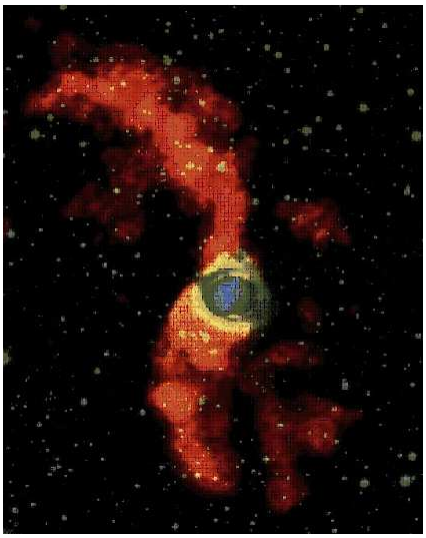

Figure 8: Overlay of three different data sets of NGC 3256. Red: neutral hydrogen spectrum. Green: the I band, representing the older stellar population. Blue: Ionized hydrogen displays star-forming regions.

3.4 What Triggers a Starburst?

The burst of star formation can be associated with interacting or merging galaxies but also with bars. In this section we will study the two possible triggers.

3.4.1 Barred Galaxies

There are several pieces of observational evidence which indicate that bars do indeed concentrate gas in the central regions of spiral galaxies, driving the gas towards the nuclear zones of galaxies.

Clear indications that nuclear starburst preferentially occur in barred host have been reported since the early eighties, from radio continuum or infrared emission and its distribution. Even though the results are not unambiguous there is a clear trend. For example, Hummel (1981) found that the central radio continuum component is typically twice as strong in barred than in no-barred galaxies.

A study of the molecular gas properties in the inner kpc of starbursts and non-starburst (Jogee et al. 2001) shows that the starbursts generally host larger central gas densities.The fact that both starbursts and non-starbursts can host a large-scale bar suggests that the lifetime of the starburst is short with respect to the timescale over which bars evolve or dissolve. The data from the study suggests that in a given barred galaxy the dominant star formation mode can change from an inefficient pre-burst phase in the early stages into a powerful circumnuclear starburst once a high enough central gas density builds up. These findings are consistent with the idea that starburst are more common in barred that in unbarred galaxies, but they also imply that the presence of a bar is not a sufficient condition for a concurrent starburst.

We conclude that statistical studies show that bars and starburst are connected, but that the results are subject to important caveats and exclusions. Further study is needed, using carefully defined samples, and exploring more direct starburst indicators, such as nuclear spectra and Hα imaging.

3.5 Interactions and Starburst Activity

Galaxy-galaxy interactions, which are a mechanism for driving reservoirs of hydrogen gas from the outer regions of the galaxies into the centre of the interacting system, may trigger starbursts. Although a recent study (Bergvall, Laurikainen, Aalto (2003)) considering two matched samples of nearby interacting and non-interacting galaxies, and measured star formation indices based on UBV colors, show that there is not significantly enhanced star forming activity among the interacting/merging galaxies. Bergvall et al.(2003) estimate from their sample that only about 0.1 percent of a magnitude limited sample of galaxies will be massive starbursts generated by interactions and mergers.

So although mergers can undoubtedly lead to massive starbursts, and most extreme starbursts show evidence for a merging/interactoin history, they appear to do so only in exceptionally rare cases. Most interactions between galaxies may not lead to any increase in the starburst activity, and those that do may be selected cases where a set of parameters, both internal to the galaxies and regarding the orbital geometry of the merger, is conducive to the occurrence of starburst activity.

4 Starburst-AGN Connection

4.1 Introduction to Active Galactic Nuclei (AGN)

An AGN has a central massive object, usually a black hole, or at least a mass concentration with total mass around 10^6-10^9 solar masses within 0.01 pc or even less, that is, about the size of the Solar System. The energy of the AGN derives from the gravitational potential energy or the surrounding material, which is released as it falls into the black hole. A considerable fraction, perhaps 10 percent, or the rest mass of this material is released in the process. The material may be gas from the interstellar medium, but it may also include stars which are disrupted as they pass within the Roche limit of the black hole. Angular momentum of the infalling material concentrates the infalling material to an accretion disc with steep gradients of angular velocity. Further collapse occurs through frictional dissipation of the differentially rotting disc, and by turbulent dissipation. Surrounding the thin accretion disc, there is a thicker torus of accreting material,which is cooler and sufficiently opaque to make the thin disc invisible from the side. The central regions can only be observed from polar directions, not from edge-on. The second added component is the pair of jets of energetic material ejected from the nucleus along the polar directions; these are astonishingly narrowly collimated, and may travel very large distances compared with the size of the torus. These jets feed energetic particles and magnetic field energy into the lobes of radio galaxies.

The galaxies must evolve in the sense that there is a stage of time prior to the onset of nuclear activity and a stage after which nuclear activity has subsided. It is known that interactions between galaxies provides the means for funnelling gas towards the central regions of galaxies, and this may feed an existing massive black hole and/or create the onset of nuclear starburst activity which could eventually lead to the formation of an AGN (Planesas, Colina and Perez-Olea 1997). Stellar bars, whose origin may be either internal or external, may also generate an inflow of gas towards the nucleus thereby leading to the onset of nuclear activity (Mulchaey and Regan 1997).

4.2 AGN Family

4.2.1 Seyfert Galaxies

Seyfert galaxies contain the most common type of nearby AGN and were the first type of active galaxy to be discovered. They are named after Carl Seyfert who in the 1940s began studying the characteristics of galaxies that appeared normal with the exception of having a very bright nucleus. At the time they were referred to as N-galaxies due to their very bright starlike nucleus (Smith 1995). They are generally spiral galaxies and range in luminosity from about 10^{36} to 10^{38} Watts, which places the brightest Seyferts as luminous as the dimmest Quasars. Seyfert galaxies have relatively high-ionisation emission lines in their spectra and are broader than the lower-ionisation emission lines observed in normal galaxies. They can be separated into two classes based on their emission line widths (Osterbrock 1989). Seyfert-1 galaxies exhibit broad emission lines, particularly hydrogen, that are attributed to ionised gas within 1 pc of the central massive black hole, whereas the spectra of Seyfert-2 galaxies only show narrow emission lines, believed to originate from a much larger region around the core. The currently favoured paradigm for unification of Seyfert galaxies asserts that the central black hole is surrounded by a dusty torus which obscures our view of the broad line region directly if viewed edge-on, resulting in a classification of a Seyfert-2 (Heisler 1998) which shows show narrower emission lines.

4.2.2 Radio Galaxies

Radio galaxies are generally associated with elliptical galaxies. They show strong radio emission well in excess of levels detected from normal galaxies, such as our own. At a typical frequency of 408 MHz, the emission is from about 10^3 to 10^7 times stronger than that of our own galaxy at that frequency (Smith 1995). The galaxy itself is not necessarily the strongest contributor of this radio emission. The source is often characterized by a double lobe structure with a central source that coincides with the optical galaxy. The lobes extend out from the central galaxy, often with a jet extending from the central source and showing hotspots near the boundary of each lobe.

Radio emission from radio galaxies is produced by synchrotron radiation as electrons move through magnetic fields. This radiation is also highly polarized due to field direction of the magnetic field. The power source for this radiation and the production of relativistic jets is again believed to be a super massive black hole. The plasma jets are believed to be produced by the winding up of magnetic fields in the accretion disk that has formed around the black hole.

4.2.3 Quasars and QSOs

The term quasar is an abbreviation of quasi-stellar radio source.They are very bright radio sources originating from what originally appeared to be an optical star-like source.the sources were located within the cores of galaxies at high red-shifts. The luminosity of these sources is in the order of 10^{40} W (some 100 times more luminous than a large elliptical galaxy in the optical spectrum) (Smith 1995). Further research has revealed other star-like sources that were radio-quiet but had similar spectra and high red-shifts. These are known as quasi-stellar objects or QSOs.

The spectra of Quasars and QSOs show a forest of absorption lines in their spectra which are believed to be due to absorption by hydrogen in the intergalactic medium

between the source and the observer. The higher the red shift, the larger the number of absorption lines.

4.2.4 Blazars

The key characteristic of blazars is that they have an almost featureless spectrum, with no absorption or emission lines from the central optically bright source and are highly variable over relatively short periods. The light from the blazer is highly polarized and has a non-thermal spectrum typical of synchrotron radiation. They radiate energetically across the spectrum from radio to gamma rays.

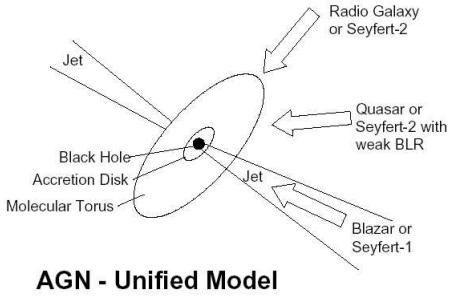

Figure 9: The Unified Model.

4.3 Starburst-AGN Connection

It is nowadays widely accepted that the energy source of active galactic nuclei (AGN) originates in the accretion of mass onto a central supermassive back hole (SMBH). However, starbursts could also play an important role in the same galaxies. In fact, many examples have been reported in which the two phenomena can coexist, and they contribute in the same fraction to the energy output. In the high luminosity regime, ultraluminous infrared galaxies (ULIRGs) are the best examples (Lutz and Tacconi 1999). However, powerful starburst are also found in less luminous AGN, such as Seyfert nuclei. Recently, the starburst-AGN connection has been suggested from the field of galaxy formation. The ubiquity of supermassive black holes in the nuclei of normal galaxies and the proportionality between the black hole and the spheroidal masses suggest that the creation of a black hole has an integral part of the formation of ellipticals and the bulge of spirals. In consequence, violent events of star formation and AGN can coexist and probably did in the past even more often that we observe today.

In the past, the role of starburst in AGN has been extensively discussed theoretically. For instance, von Linden et al. (1993), suggest that starburst-induced turbulence in the ISM is responsible for the final stages of infall of gas into the nuclear engine. Weedman (1983) pioneered an AGN model involving a large number of small accretors - the remnants of the massive stars in a compact circumnuclear starburst, preceding the AGN stage. It has also been proposed that a starburst can mimic an AGN (Terlevich and Melnick, 1985). Terlevich et al. (1992) showed that at least low luminosity Seyfert 2s can be explained via active circumnuclear star formation alone. Still another possibility is that circumnuclear starbursts strongly alter the observed properties of AGNs. For example, Ohsuga and Umemura (1999) consider a model in which the wind from the circumnuclear starburst forms an obscuring dust wall. The alternative process, in which an AGN disturbs the

galactic ISM and triggers a starburst was reported by van Breugel et al. (1985), van Breugel and Dey (1993), and Dey et al. (1997). This behavior is usually related to the expansion of a radio source into the ambient medium. Also, once formed the black hole, can gravitationally disturb the interstellar gas, triggering a starburst.

A better understanding of the frequency of AGN and starburst coincidences is required to probe the connection between the two phenomena. A conventional method to search for young starbursts is to look for stellar absorption features in the ultraviolet spectra of galaxies (e.g. Heckman et al., 1997; Gonzalez-Delgado et al., 1998). Current instrumentation allows this approach to be applied in just a few cases; Heckman (1999) found only a few number of Seyfert 2 nuclei bright enough to obtain UV spectra with sufficiently high signal-to-noise to identify young stars. It is possible that they show starburst signatures because of the selection bias toward galaxies with high surface brightness, which could be a result of the active star formation. In addition, application of this technique to AGNs is restricted by the extinction. Since the ISM lies around the AGN in a complex three-dimensional distribution, a significant component of the observed UV flux may come from the stellar population in front of the active nucleus and not be associated closely with it.

The infrared range has an advantage in probing AGN environments because of the lower extinction. Since the strengths of the CO bands at 1.6m and 2.3m, they can indicate the presence of young red supergiants. The CO band strength is also a function of metallicity. However, in the bulges of large spirals, the typical haunt of a Seyfert nucleus, the surface gravity dependence dominates. Empirically, quiescent spirals have a very narrow range of 2.3m band strengths (Frogel et al., 1978; Frogel, 1985) and most starburst galaxies have bands substantially deeper than this value. Although the 2.3m strength has proven useful in identifying and studying starbursts, in applying similar techniques to Seyfert galaxies one must contend with the possibility that the band can be diluted by radiation from the warm dust surrounding the central engine in the AGN. Oliva et al. (1995) therefore used the second overtone at 1.6m to probe the environments of AGNs. Unfortunately, this band is weak and its dependence on metallicity is strong comparable to 2.3m CO band. To avoid these issues, Oliva et al. (1999) derived the mass to luminosity of the nuclear region and identified starburst/Seyfert combinations where this ratio was low, rather than relying on the CO band strengths. Both of these approaches require high signal to noise spectra of moderately high resolution. Thus, they can only be applied to AGNs that are bright in the near infrared, which may result in a selection bias toward examples with strong circumnuclear starbursts.

4.3.1 Typical Evolution

When a late-type galaxy undergoes a transient encounter, gas is supplied from the disk to the circumnuclear region at a high rate and induces a luminous starburst. Then the gas is transported into the nuclear region. The starburst galaxy is expected to become a starburst- dominant Seyfert galaxy if there is the central BH.

The duration of gas supply is 10^8 yr. Thereafter the starburst and AGN become less luminous. The object is expected to become a host-dominant Seyfert galaxy before it becomes an inactive galaxy. The object does not become an AGN-dominant Seyfert galaxy because the central BH is not massive enough Throughout the evolution, the object remains a late-type galaxy. More time is required for the morphological type to change markedly. A similar evolution is expected when a late-type galaxy undergoes a minor merger.

When an early-type galaxy undergoes a transient encounter or a minor merger, the

rate of gas supply is not high. The object is expected to become an AGN- Seyfert galaxy if there is the central BH. The AGN-dominant Seyfert galaxy is expected to become a host-dominant Seyfert galaxy.

When a galaxy undergoes a major merger, the rate of gas supply is high if the amount of available gas is not very small. The object is expected to become a starburst galaxy $\rightarrow$ a starburst-dominant Seyfert galaxy if there is the central BH. Since the duration is as long as 10^9 yr, all the gas within the galaxy is transported into the circumnuclear and nuclear regions and is consumed by the starburst and AGN. The accretion onto the BH increases its mass and hence the possible maximum luminosity of the AGN. Probably via an AGN-dominant stage, the object is expected to become a host-dominant Seyfert galaxy.

Thus, while a high rate of gas supply is necessary for a starburst, a low rate of gas supply is sufficient for an AGN. Only when the rate of gas supply is high, the luminosity of a circumnuclear starburst dominates over that of the AGN.

The rate of gas supply is determined by two parameters: (1) the amount of available gas within the galaxy and (2) the efficiency for transporting the gas into the circumnuclear or nuclear region. The first parameter is related to the morphological type of the galaxy and the second to the mechanism that generates a distortion and thereby induces gas inflow. Various mechanisms with various efficiencies are available, we will center in galaxy interactions.

The amount of gas and the efficiency for transporting the gas put serious constraints on the luminosity of a starburst, which requires a high rate of gas supply. The constraints are not serious to the luminosity of an AGN which does not require a high rate of gas supply.

To sustain the luminosity, a starburst requires a higher rate of gas supply that an AGN. The rate of gas supply is determined by the amount of available gas within the galaxy and the efficiency for transpoirting the gas into the circumnuclear or nuclear region. The amount of gas is larger in a later-type galaxy. Since the central BH is less massive, the possible maximum luminosity of the AGN is lower. Starburst-dominant Seyfert galaxies are of later types than AGN-dominant Seyfert galaxies.

5 Conclusion

Starbust activity can be driven by bars or interaction between galaxies, although according to recent statistical studies most of the interaction may not lead to any increase in the starbust activity. Moreover, the intense star formation and nuclear activity often go hand in hand. There is a clear relation between AGN and starburst, but many questions are still unanswered. Hopefully the new generation of ground-based telescopes will decipher the nature of the starbust-AGN connection in proto-galaxies. In the meantime much effort should be invested in predicting what we should expect to see.

Acknowledgements

I would like to thank Alessandro Rome who put so much effort in this course and encouraged us so much. Also I would like thank all my fellow students who help me to understand new concepts in astrophysics with their extraordinary job.

References

Bergvall N., Laurikainen E., Aalto S., 2003, A&A, 405, 31

Dey A., Breugel W., Vacca W.D., Antonucci R., 1997, ApJ, 490, 698

Doyon R. Joseph R.D. Wright G.S. 1994, ApJ, 421, 101.

Frogel J.A., Persson S.E., Aaronson M., Matthews K., 1978, ApJ, 220, 75

Frogel J. A., 1985, ApJ, 298, 528

Gonzalez-Delgado R.M., Heckman T., Leitherer C., Meurer, G., Krolik, J., Wilson, A.S., Kinney, A., Koratkar, A., 1998, ApJ, 505, 174

Graham J.R., Wright G.S., Meikle W.P.S., Joseph R. D., Bode, M.F. 1984, Nature, 310, 213

Heckman T.M., Gonzalez-Delgado R., Leitherer C., Meurer G.R., Krolik J., Wilson A.S., Koratkar A., Kinney A., 1997, ApJ, 482, 114

Heckman T.M., 1999, preprint (astro-ph/9912029)

Heisler C.A. (1998), Orientation and Evolutionary Effects in Active Galactic Nuclei, Publ. Astron.Soc. Aust, 1998, 15, 167-75.

Hummel E. 1981, A&A, 93,93

Jogee S. 2001, in Starburst Galaxies: Near and Far, eds. L. Tacconi D. ,Lutz (Heidelberg: Springer-Verlag),182

Kinney A., Bohlin R.C., Calzetti D., Panagia N., Wyse R. F. G., 1993, ApJS, 86,5

Moorwood A. F. M., Oliva E., 1994, ApJ,429,602

Mulchaey J. S., Regan M. W., ApJ 482, L135 (1997)

Norris R.P., Forbes D. A., 1995, ApJ, 446, 594

Ohsuga K., Umemura M., 1999, ApJ, 521, L13

Oliva E., Origlia L., Kotilainen J.K., Moorwood A.F.M., 1995, A&A, 301, 55

Oliva E., Origlia L., Maiolino R., Moorwood A.F.M., 1999, A&A, 350, 9

Osterbrock D. E., (1989), Astrophysics of Gaseous Nebulae and Active Galactic Nuclei, University Science Books

Peng R., Zhou S., Whiteoak J.B., Lo K.Y., Sutton E.C., (1996). ApJ, 470, 821.

Planesas P., Colina L., Perez-Olea D., (1997) A&A, 325,81

Shen J. ,Lo. K.Y., 1995, ApJ, 445, L99

Smith E. P., Kassim N.E., 1993, AJ, 105, 46

Smith R. C., (1995), Observational Astrophysics, Cambridge University Press

Terlevich R., Melnick J., 1985, MNRAS, 213, 841

Terlevich R., Tenorio-Tagle G., Franco J., Melnick J., 1992, MNRAS, 255, 713

van Breugel W., Filippenko A.V., Heckman T., Miley G., 1985, ApJ, 293, 83

van Breugel W.J.M., Dey, A., 1993, ApJ, 414, 563

von Linden S., Biermann P.L., Duschl W.J., Lesch H., Schmutzler T., 1993, A&A, 280, 468

Weedman D.W., 1983, ApJ, 266, 479

Extrasolar Planets

Markus Janson

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(f00maja@dd.chalmers.se)

*

Abstract

Through innovative methods and developments in technology, extrasolar planets were detected for the first time starting about 10 years ago and have continued to be found at a rapid pace since then. More and more is discovered about individual giant planets, but the quantitative observational background needed for a solid theory behind the formation of planetary systems is still lacking. In this paper, the present theoretical status of planetary systems is described as well as the techniques used for detections and planned improvements for the future. It is concluded that an exciting future can be anticipated in the field of planetary science.

1 Introduction

The very question of whether extrasolar planets exist is necessarily preceded by two revelations. The first one is the heliocentric paradigm, first suggested by Copernicus in about 1507, which states that Earth and the other planets revolve around the Sun. The second is the realisation that the Sun is the very same kind of object as the stars observed all over the sky. Knowing these two things, the question of whether these stars also have planetary systems and how they in that case look is obvious. The answer, however, is difficult to provide, because of the huge interstellar distances and the fact that planets do not radiate to nearly the same extent as their harboring stars. Only for about the past ten years, through clever methods and technological development, have we been able to start investigating this issue. To date, 120 extrasolar planets have been discovered around main sequence stars in our stellar neighbourhood. In this paper, a review of extrasolar planet research will be presented. The current theories regarding planetary formation will be investigated, followed by a description of the methods used for detecting extrasolar planets and their current and future abilities. Thereafter it will be described what these observations have shown and how this affects the formation scenarios. Finally there will be a section dealing with a kind of planet still impossible to detect, but of all the more interest: Earth-like planets.

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

2 Planetary Formation

To fully understand why planets form around stars, we need to start by looking at how the star itself forms. What is to become one or several stars, starts off as a dense cloud of particles in space. The constituents of such a cloud are hydrogen (to a great majority), a bit of Helium and a small fraction of heavier elements (all grouped together as "metals") – to which extent depends a bit on the local conditions, but mainly on age (because metals originate exclusively from earlier generations of stars).

If this cloud cools at a sufficiently rapid rate, it will fragment, and the fragments will collapse towards the centre of their respective masses. In addition, each fragment in general has an internal angular momentum which it conserves during the process, which causes it to flatten – that is, it collapses towards a plane which is perpendicular to the axis of rotation. This way, at the end of the collapse each former cloud fragment consists of a central massive bulge, which will become a star, and a circular, planar disc around it, which will provide the materials needed to form planets. Some of the matter from the disc falls on to the central object which at some point, if it is massive enough, will ignite hydrogen burning and start radiating like is to be expected from a normal (main sequence) star. Through redistribution of angular momentum, the disc will eventually stabilize. Also, for some time after the collapse, remaining gas from the cloud surrounding the newly born solar system will fall on to the disc, which is why this disc is generally referred to as the accretion disc.

This far, the theory is quite uncontroversial. Star forming is known to occur in dense, interstellar clouds, and in numerous cases, accretion discs (which are large and moderately dense and thus relatively easy to observe) have been observed around young stars. However, from this point on it becomes much more uncertain. This is largely due to observational limitations. As we shall see later, observation of planets is a tricky business – and as a result, the only planetary system we have a reasonably complete picture of is our own. This is not necessarily a good representative of planetary systems in general, and as it has turned out, certainly does not provide a unique representation of how planets can be distributed in the system.

Still, in particular for terrestrial planets, there is one generally favoured theory for the rest of the planetary formation process, which typically (and without exception in this essay) is referred to as the collisional accumulation model. For giant planets, it is slightly more complicated. There are two prevalent theories; the still favoured (although increasingly challenged) theory for giant planet formation is the core accretion model. The most plausible alternative scenario is the model of disc instability. Both theories and an appropriate discussion will be included below, following a description of collisional accumulation.

2.1 Collisional Accumulation

The model in its basics is as follows: Small particles in the disc orbiting the star begin to stick together, forming slightly larger structures hierarchically. After some time, the matter is generally concentrated to grivels and rocks uniformly distributed in a disc around the star, almost equivalent to the system of Saturn and its rings. From here, the rocks continue to grow together through gravitational interaction until they reach a stage where they are reffered to as "planetesimals" with sizes on the order of a kilometer. At this point the system is more like an equivalent to a central star and different asteroid belts at various distances from the star with relatively empty spaces between them.

The planetesimals of a certain "belt" will of course keep gravitationally interact with each other (with increasing strength as they grow larger), alter each others orbits and occasionally collide to either form a larger object or shatter (if the velocity of the collision is too high). Depending on how much material is available at and near the distance of the soon-to-be planet, the hierarchical forming of larger structure can stop at the size on the order of Mercury to Earth, or if there's more material to go around, keep growing for yet another while.

2.2 Core Accretion

This point up to which we have outlined the process so far is possibly where core accretion starts taking place as a continuation of core accumulation. Once (if) the planetesimal reaches a certain critical mass, it will start accreting gas at a high pace (in other words, gas is accreted on to a solid core, which is where the name "core accretion" comes from), and eventually a gas giant is formed.

This model describes our own planetary system quite well – near the sun, where the accretion disc should have been relatively sparse, small unaccreted planets occur, and further out, where the disc should have been denser, gas giants are situated. However, there is a problem of timescales: in the standard core accretion scenario, the gas giants take too long to form – longer than the estimated lifetime of the accretion disc. For this reason, another theory is also widely considered: the model of disc instabilities.

2.3 Disc Instabilites

In this scenario, planets are formed in a very similar way to how stars are formed as described earlier. The disc is cooled and as a result becomes gravitationally unstable (from which the name of the model originates); it fragments, and each fragment collapses towards the center of its mass until it is dense (and hot) enough so that the collapse ends, at which point we have a planet.

As has been implied, this model would solve the problem of timescales in giant planet formation – giants are formed here over the course of on the order of only thousands of years in simulations. However, this model also has its fair share of problems. For instance, it may be necessary for instabilities in an accretion disc to be triggered rather than them occuring spontaneously in the disc. Also, heating processes in the disc could delay or completely stop the process of gravitational instabilities in the cloud.

2.4 Discussion of Theories

Both theories (core accretion and disc instabilities), as has been stated, have their respective advantages and problems. A recent study by Rice and Armitage (2003) demonstrates a possible solution to the core accretion time scale problem. By considering possible turbulent behaviour of materials in the disc, protoplanets are in Rice and Armitages simulations allowed to random walk radially through the disc, rather than remaining at a static radial distance from the star during the entire formation process (like in the original core accretion model). This way, during the growth of the core it has a much larger cross-sectional area for receiving new material to make it grow larger. In the original core accretion model, the static core would gradually deplete the abundance of planetesimals in an area around it, causing less collisions and thus constraining the rate of growth towards the critical mass. With random walk considered, the area becomes bigger and thus

the depletion rate is slower. This makes the core able to reach critical mass for runaway gas accretion at a much faster pace.

The gravitational instability model problems are also not without possible solutions. Numerical simulations have been performed by e.g. Boss (2002a) under various thermodynamical conditions in the disc which show that fragmentation of a gravitationally unstable disc can occur under certain of these conditions. Boss also puts forth the idea that extreme UV photoionization could cause evaporation in the outer parts of the disc such that the planets residing there would lose large parts of their gas envelope and end up as ice giants rather than gas giants such as Uranus and Neptune – these would otherwise constitute a problem for the gravitational instability theory.

Even so, problems and unclarities remain in both theories. The reason for this situation is mainly because observational input is simply inadequate. As was mentioned earlier, the only planetary system we have a decent view of is our own, and even here, vital information that would be needed for a complete picture is missing such as initial thermodynamical conditions and the properties of gas dynamics in the disc. In order to be able to differentiate the theories, detailed high-resolution studies of accretion discs must be performed. Also, a more statistically complete view of extrasolar planetary systems must be achieved. Projects designed for dealing with these issues are underway, and are discussed in the "Extrasolar planet observations" section.

An example of a relevant observational quest for within the solar system is to determine the core masses of the gas giants – Jupiter and Saturn. Recent studies on the subject were able to set an upper limit for the masses of the respective cores. These limits are lower than the prediction of core mass achieved from the core accretion model (and allows the core mass even to be inexistant), which has been said to speak in favour of disc instabilities (see e.g. Mayer et al. 2003; Boss 2002a). However, the already mentioned simulations by Rice and Armitage create a way around this since they introduce the possibility of a core mass below the measured upper limit.

If the mass of the core of Jupiter should be found to be zero (which is unlikely, yet possible), this would hugely speak in favour of the disc instability model since the core accretion model demands a core for the gas to be accreted on, whereas disc instability has no such requirements. Note, though, that if Jupiter should be found to have a non-zero mass, this does not necessarily disqualify the disc instability model, because there is still a possibility that solid material trapped inside the collapsing disc fragment could accumulate and create a core after the planet has formed, although it is not clear how such a process would occur (see Boss 2002b).

3 Extrasolar Planet Observations

Detecting extrasolar planets is a difficult task and therefore a young field of reasearch (in the sense of actual achievement, anyway) – roughly ten years to date. The knowledge of this subject is, while currently modest, rapidly growing and new exciting projects and methods are proposed and developed continuosly along with technological progress. In this section, there will follow a description and discussion of the major tested or suggested methods for detecting planets outside of our own solar system. Examples of major projects (in progress or planned) will be mentioned for the respective categories in the event that such are in existance.

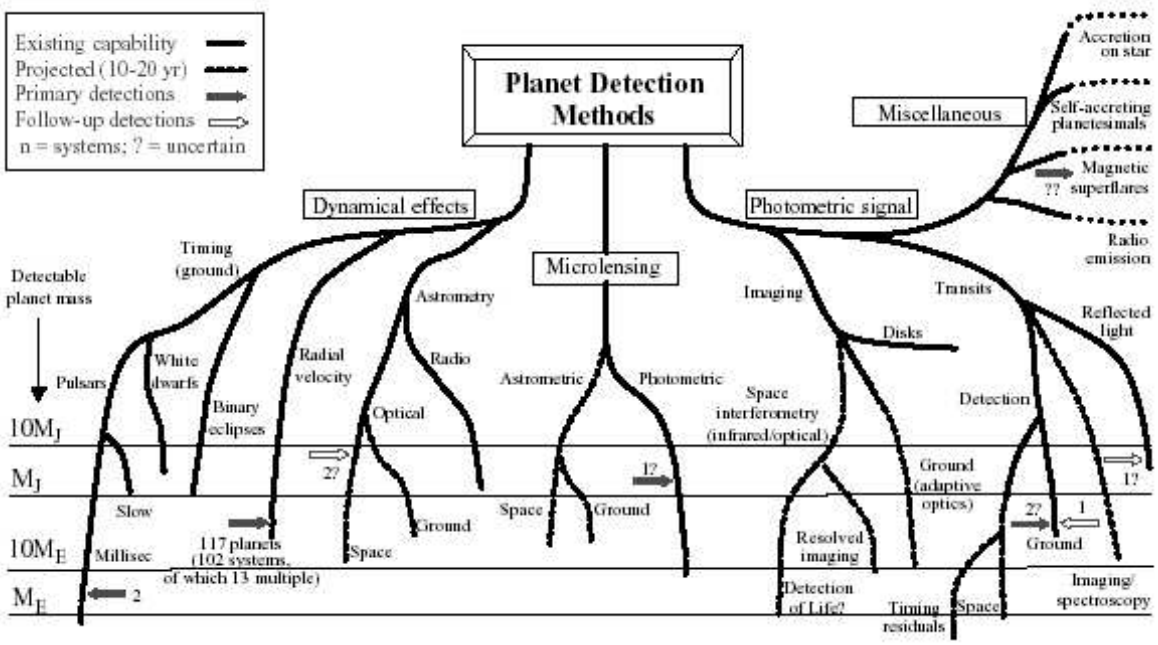

Figure 1: The Perryman tree, originally appearing in Perryman (2000). This is an updated version from Schneider (2004a).

3.1 Radial Velocity

We're starting off with the absolutely most succesful detection method in terms of number of detections – radial velocity. The first extrasolar planetary companion to a sun-like star was detected using this very method by Mayor and Queloz (1995). Since then, more than 100 planets have been detected with radial velocity measurements. Like some other detection methods (astrometry and pulsar timing), radial velocity takes advantage of the gravitational effect a planet has on its host star. Keplerian motion in celestial mechanics is often simplified only to consider the motion of a less massive object around a more massive one, but in fact, both bodies circulate around their mutual centre of mass. Obviously, in the case of a star-planet system, the centre of mass is much closer to the star and thus the star moves extremely little compared to the planet. Still, this stellar motion is measurable with present techniques under certain circumstances.

In the case of radial velocity, the component of the stellar velocity which is parallell to the line of sight (thus, of course, the name) is studied through the Doppler shift of its spectrum. If the star e.g. has one planet in an edge-on circular orbit as seen from earth, it will istelf move in a circular edge-on orbit according to the above reasoning, which produces a sinusoidal velocity variation in time along the line of sight that we can measure with said technique.

In the general case, the radial velocity amplitude K is given by

$$K = (\frac{2\pi G}{P})^{1/3} \frac{M_{\rm p} \sin i}{(M_{\rm p} + M_{\rm star})^{2/3}} \frac{1}{(1-e^2)^{1/2}} \quad (1)$$

where P is the period of the orbit, G is the gravitational constant, $M_{\rm star}$ is the mass of the star, $M_{\rm p}$ is the mass of the planet, i is the inclination of the orbit in the sense of angular deviation from a face-on orbit as seen from earth and e is the eccentricity. Taking several measurements of the radial velocity of a star with a planet and making a best fit

to this data, one can decide the velocity amplitude from the amplitude of the resulting curve, the period from the periodicity and the eccentricity from the specific shape of the curve (see figure 2). Since $M_{\mathrm{star}} \gg M_{\mathrm{p}}$, $M_{\mathrm{p}} + M_{\mathrm{star}}$ in formula (1) can roughly be replaced by M_{star}, which can be estimated from e.g. the spectral type of the star. What remains then is only $M_{\mathrm{p}} \sin i$.

From this we can draw two major conclusions. The first one is that if we can isolate a periodic feature in a radial velocity versus time diagram of a star, we can decide a lower limit on the mass of the planet responsible for the movement; this mass estimate is exact if the inclination is 90 degrees, but higher if it is not. The second conclusion is that higher mass planets cause a larger velocity amplitude, and the instruments used for detection have a limited sensitivity due to noise. This means that more massive planets are easier to detect, and so we will get a measurement bias towards high-mass planets. Also, note that planets at or close to zero inclination would be impossible to detect by this method.

In addition, Keplerian motion implies that $P \propto a^{3/2}$ where a is the semi-major axis of the planetary orbit. Thus we can uniquely determine the distance from the planet to the star by the above method. We can also see from this proportionality and equation (1) that planets with smaller semi-major axis are easier to detect with the radial velocity method. This bias is further amplified by the fact that planets with large semi-major axis (and thus longer periodicity) must be monitored for a longer time in order to get a clear detection. A typical so-called "hot Jupiter" (see the "Extrasolar planet properties" section) takes only a few days to complete a full cycle giving a sinusoidal pattern needed to confirm the presence of a planet. A planet at the same distance from its host star as Jupiter's distance from the sun, however, would demand twelve years of monitoring in order to produce an equally reliable signal, even if we'd assume it to be so massive that it would create the same velocity amplitude of the star as the close-in hot Jupiter.

As a final note on the principle of radial velocity, the formula for the velocity amplitude in itself contains no dependence on the distance between Earth and the planet-hosting star, but this does not mean that we can detect planets around stars with this method completely independently of their distance. This is because the possibility of distinguishing a radial velocity component from noise depends on the apparent brightness of a star, which of course in turn depends on the distance to it (in addition to its actual brightness, naturally). This effectively gives a detection possibility dependence on distance, although not to such a dramatic extent as we shall see is the case for e.g. astrometry.

Radial velocity is the extrasolar planet finding technique with the most projects dedicated to it, since it is the currently most efficient method available, and in difference from some other methods is relatively insensitive to atmospherical effects and therefore possible to implement with full efficiency using ground-based equipment. Two of the more succesful projects in this area are ELODIE (see e.g. Mayor et al. 2004) and CORALIE (see e.g. Naef et al. 2004) which are spectrographic telescopes monitoring selected stars in the northern and southern hemispheres, respectively. These instruments typically reach accuracies of around 10 m/s (ELODIE somewhat above, CORALIE somewhat below; the exact accuracy depends on the properties of the star, as mentioned), which is just enough for detection of Jupiter-equivalent planets in edge-on orbits. For this reason, detections are of course generally either more massive or at a closer orbit than Jupiter (see figure 3). Most other radial velocity projects have accuracies similar to this or in some cases even somewhat better. The best methods currently under development are thought to ideally be able to produce accuracies of about 1 m/s. This would be more than adequate to detect planets equivalent to Saturn, but not Uranus, and still very far off from detecting Earth equivalents.

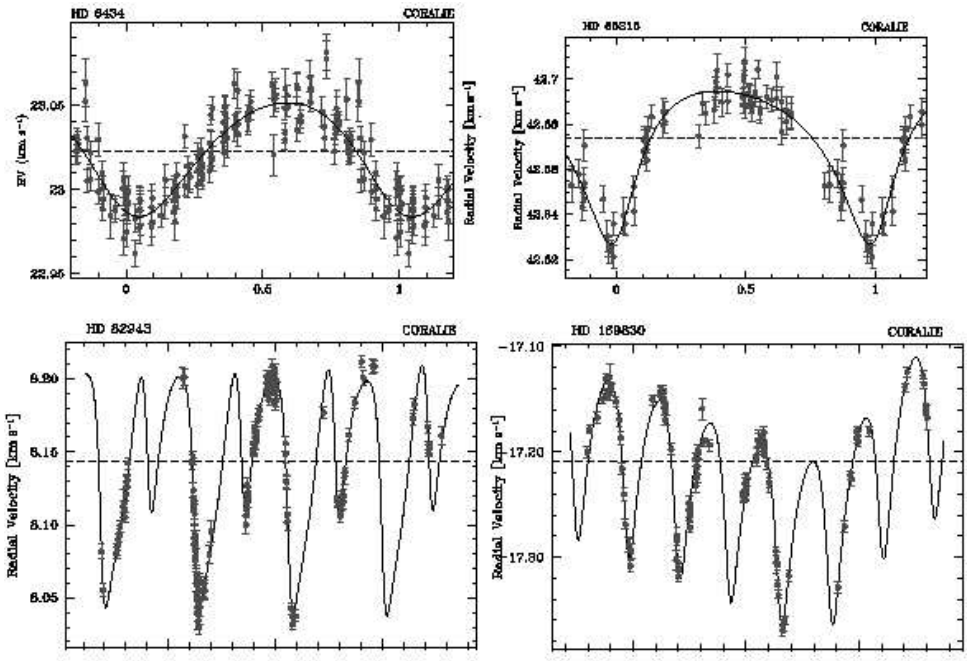

Figure 2: Examples of radial velocity curves. Upper left: A nearly circular orbit. Upper right: An elliptical orbit. Lower left: A resonant two-planet system. Lower right: A non-resonant two-planet system (from Mayor et al. 2004).

3.2 Astrometry

Just like for radial velocity, astrometry entails observing the motion of a star caused by the gravitational perturbation from a planetary companion, but whereas in radial velocity searches, the movement parallell to the line of sight is studied, astrometry means studying the movement in a plane perpendicular to the line of sight.

A star encircled by a planet generally exhibits an elliptical motion on the sky. The angular semi-major axis of this ellipse is given by:

$$\alpha = \frac{M_{\mathrm{p}}}{M_{\mathrm{star}}} \frac{a}{d} \tag{2}$$

where a is the semi-major axis of the planet's motion and d is the distance between the star and the observer. That is, if we can measure the angular semi-major axis from this particular component of the movement of the star on the sky, estimate the mass of the star e.g. from its spectral type, determine the distance e.g. through parallax data and calculate semi-major axis of the planet from the periodicity of the movement, then we can determine the mass of the planet. This is an obvious advantage over the radial velocity method, which as mentioned only gives a lower limit on the mass. However, astrometry has several downsides.

First of all, as is seen from formula (2), the method is sensitive to distance – the further out you go, the better resolution (in direct proportion) you need in order to detect an equivalent planetary system. Secondly, the resolution is a problem already for the very shortest interstellar distances of a few parsecs. Atmospherical effects necessarily limit the possible resolution to about 1 mas for ground based telescopes, regardless of the capacity of the instrument. This means that quantitative astrometry would have to be performed by using space-based telescopes, which of course is more expensive and complicated.

Also from equation (2), we can see that angular size of the star's movement is proportional to the semi-major axis of the planetary orbit. This is due to the fact that when a planet of a certain mass is further away from the star, the centre of mass of the system is further away from the center of the star, which causes a larger axis for the rotation of the star. This means that the astrometry method is increasingly efficient at larger semi-major axes (although the problem of increasing period with increasing semi-major axis as discussed in the radial velocity section of course remains). This creates a specific area of interest for astrometry in the context of extrasolar planet detecion as a whole, since most other methods are biased towards small orbits.

As has been mentioned, astrometry capacity from the ground is limited, and as such it is not surprising that the recent first certain detection of an extrasolar planet through astrometry was achieved by the space-based Hubble space telescope (see Benedict et al. 2002). The planet, revolving around the star Gl 876 (which also has another planet), was first found using the radial velocity technique, and astrometry was then used to determine the mass ($1.89 \pm 0.34 M_J$) and inclination (84 ± 6 deg – that is, nearly edge on) of the planet.

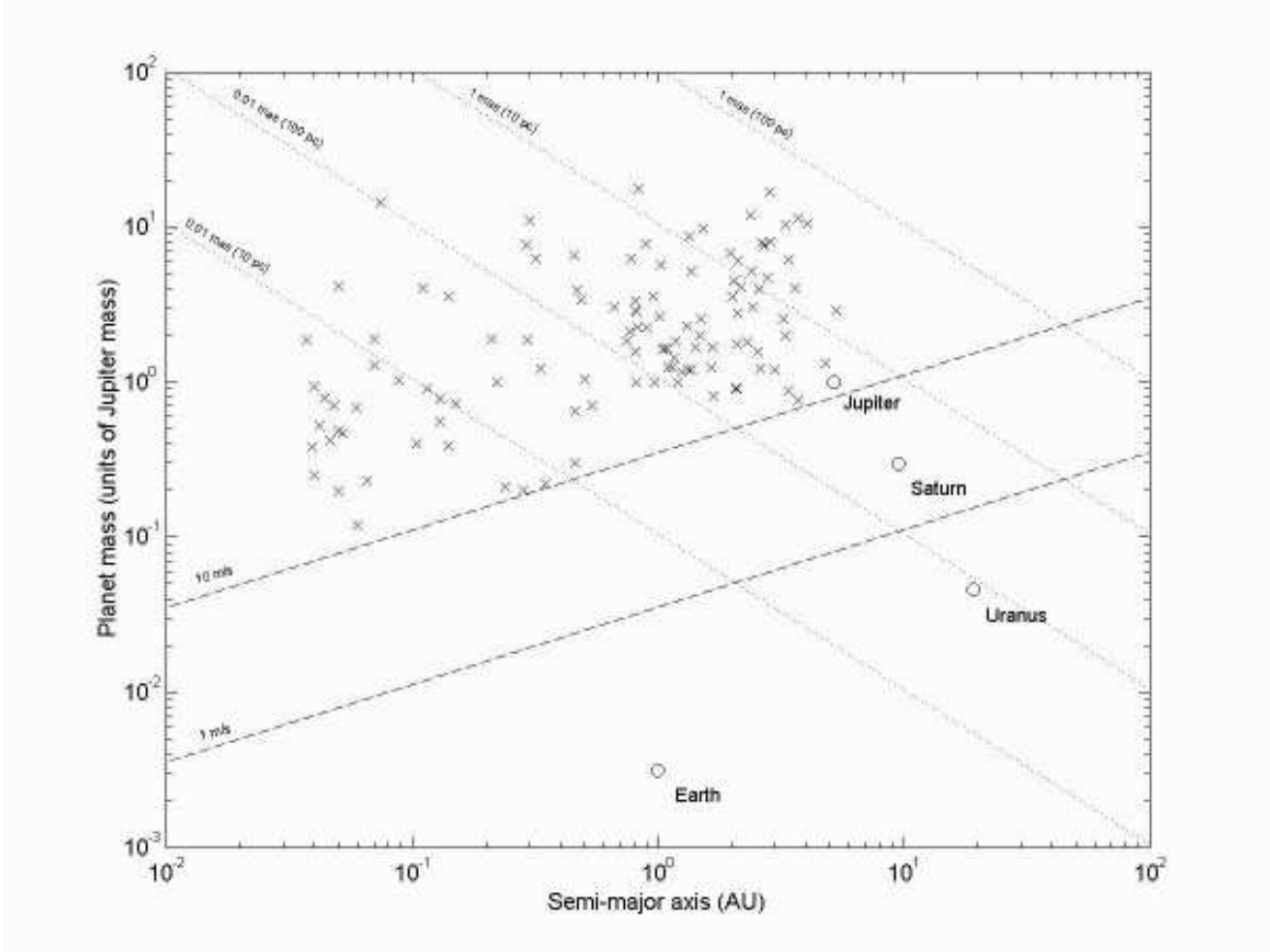

Figure 3: The dashed lines represent detection limits for radial velocity with different precision, and the dotted lines represent detection limits for astrometry. Figure concept from Perryman (2000). Figure recreated and updated by myself using data from Udry (2004).

While atmospheric effects as discussed limit the resolution of observations, ground based astrometry is not entirely hopeless. By using intricate ways of compensating for these effects, one could reach accuracies a bit below 1 mas with interferometric telescope sets such as the VLTI (Very Large Telescope Interferometer). Still, in summation, the main future application for astrometry is most likely space-based searches for massive planets on large orbits (such as Saturn and Uranus equivalents) and determination of mass and inclination for systems already found by radial velocity measurements in appropriate mass and orbit size ranges.

3.3 Transits

When a planet passes between a star and the observer, some of the light from the star is obviously blocked as seen from the observer. If the luminosity difference cause by this event is large enough, such a planet can be detected and from the properties of the luminosity dip and periodicity of events, conclusions can be drawn about the properties of the planet. This is known as the transit method.

The luminosity difference is given by

$$\frac{\delta L}{L_{\mathrm{star}}} = (\frac{R_{\mathrm{p}}}{R_{\mathrm{star}}}) \tag{3}$$

where R_{p} and R_{star} are of course the radii of the planet and star, respectively. That is, the method favours detection of large planets. There is however much more to transit detections than what can be read out from this formula. First of all, the orbit has to be at or very close to an edge-on orbit towards the observer, otherwise it will never pass in front of the star. Also, we prefferably want to measure more than one transit, to be sure that it really is a satellite event and determine its periodicity (and thus semi-major axis). This contributes to that small orbits are preffered for transit detection. Aside from that a longer period means that a star has to be observed for a longer period of time before a transit can be found in the first place, it also means that once you've found such an event you would have to wait for maybe several years (12 years in the case of a Jupiter equivalent) until the next one.

The strict requirements on inclination of course limits the amount of cases suitable for transit analysis drastically. That the method is still very interesting and common in scientific literature is that once applicable, it opens up several very interesting possibilities in combination with radial velocity measurements (which also favour near edge on orbits for detection). From radial velocity, $M_{\mathrm{p}} \sin i$ can be determined, and from transits we can find the inclination, which enables us to deduce the mass of the planet. Also from transits, we know the radius of the planet which makes it possible to calculate the bulk density.

The list of applications for the transit method doesn't end at determining orbital and physical properties of the celestial bodies in themselves; by comparing the spectral properties of the starlight while no transit is occuring with those appearing during a transit, conclusions can be drawn about the atmosphere of the planet (some of the light during a transit passes through the planetary atmosphere and gives rise to absorption features). This has been done on the first of four planets so far detected: HD 209458 b. This led first to the detection of sodium, and then to what is interpreted as escape of atomic hydrogen from the planetary atmosphere (see Vidal-Madjar et al. 2003) through absorption in the Lyman α line, and recently also to the detection of carbon and oxygen (see Vidal-Madjar et al. 2004). Also, of course, properties such as mass and size were

deduced in the way described above. These measurements confirm that HD 209458 b is indeed a gas giant planet, rather than for instance an invisible binary component (such as a brown dwarf) in a low-inclination orbit as has been suggested as an alternative interpretation to some of the planets found through radial velocity.

HD 209458 was first detected through radial velocity with the Keck spectrograph, after which it was targeted (being a hot Jupiter and thus probable candidate) for a transit search using STARE, a photometric instrument specifically designed for transit search in a large sample of stars. The transit was found in 1999 (see e.g. Charbonneau et al. 2000). Another known transiting planet, OGLE-TR-56, was found in 2002 through transits in the OGLE search, which is primarily a project for detecting microlensing events (see the "gravitational microlensing" section). It was later confirmed through radial velocity with Keck (see Konacki et al. 2003). Since then, two more confirmed planets have been detected with OGLE.

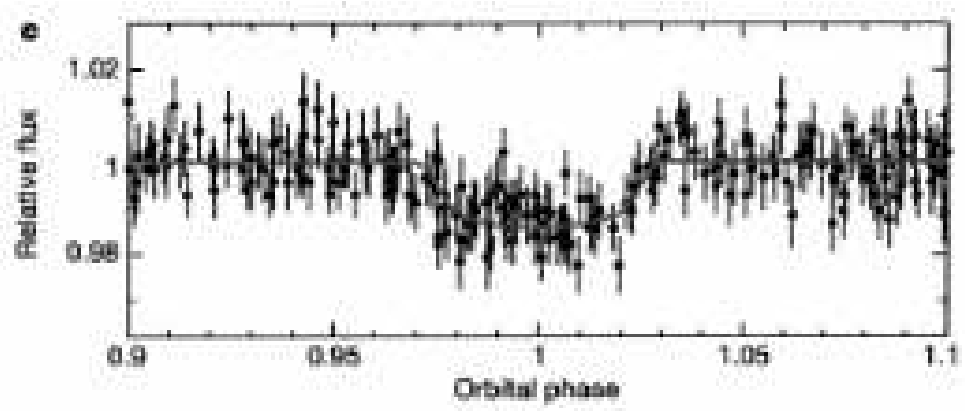

Figure 4: Transit curve of OGLE-TR-56 (from Konacki et al. 2003).

It is even possible to detect moons around extrasolar planets with the transit method. Such an object would have two effects that could show up in the transit data; it could block out additional light from the star at certain times (when it is not aligned with its parent planet along our line of sight) and also, it would cause the transiting planet to wobble which could make some transits begin earlier or finish later compared to others. In Brown et al. (2001), such an investigation is described using transit data of HD 209458 from the Hubble space telescope. No satellite was found. It was concluded that in order to find a satellite in this case, it would have to either have a size of more than $1.2R_{\text{Earth}}$ (in order for the darkening effect to show up) or a mass of more than $3M_{\text{Earth}}$ (for the wobble effect to show). With a more precise instrument, these limits can be reduced further. In the particular case of HD 209458, it has been estimated (see Barnes & OBrien, 2002) that since it is so close to its star, it could only keep a satellite in a stable orbit if the satellite mass is less than $7 * 10^{-7} M_{\text{Earth}}$ which would most certainly not show up in transit data. However, the allowed satellite mass for stable orbits increases with increasing distance between planet and star. Earth mass satellites are allowed at about 0.3 AU, and outside of 0.6 AU there is no meaningful mass constraint in this sense. In conclusion, this means that a search for moons around extrasolar planets is meaningful for transitting planetary orbits at about 0.5 AU and upwards with present techniques.

The role of transits in the future will most likely continue to be (relatively) high quality measurements of a few select systems, due to its constraint of inclination. The only situation in which transits could be appropriate for large scale detections is when a large

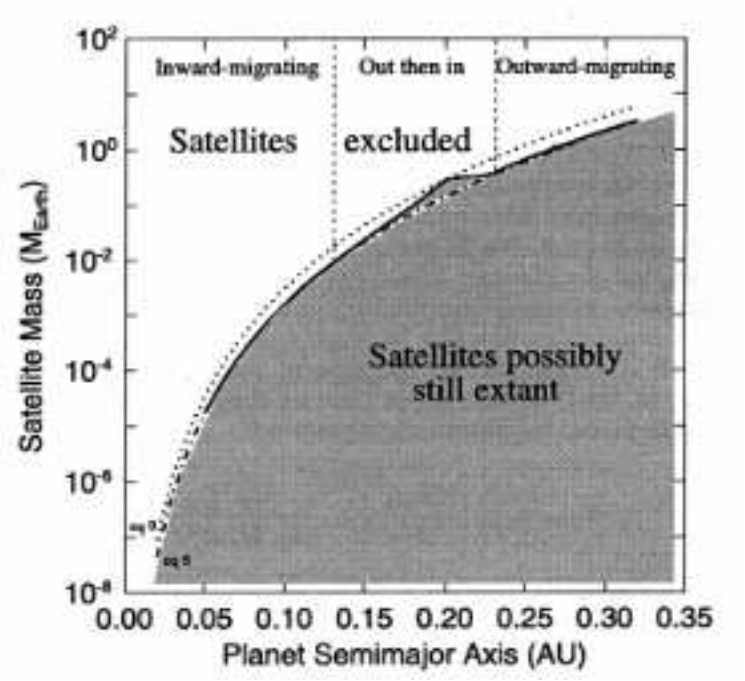

Figure 5: Theoretical mass constraint on hot Jupiter satellites (from Barnes & O'Brien 2002).

amount of stars are collected in a small area and thus can be investigated simultaneously, such as in a cluster. Such a survey has been performed on the 47 Tucanae cluster with the Hubble space telescope, which statistically should have yielded several detections of hot Jupiters, judging from the frequency at which such have occurred in radial velocity searches. However, none were found. This will be discussed in the "Metallicity" section.

3.4 Gravitational Microlensing

Like for radial velocity and astrometry, planet finding through gravitational microlensing makes use of the planet's gravitational properties, but whereas the two former ones study the effect from the planet on its host star, the latter studies the effect from the planet on light from a background star.

Similarly to how the trajectories of massive objects are bent in the presence of a gravitational field, light is also bent by gravitation. Thus when a massive object in space passes between an observer and a luminous background object, the middle object bends the light from the source and becomes like a lens, collecting a larger amount of light towards the observer than is normally reached from the background source. This is the principle of gravitational microlensing. In the context of planet finding, the background lightsource is prefferably a star in e.g. the galactic centre bulge, and the lens is for instance a star with a planetary system which is passing in front of the lightsource.

The magnification (increase of received flux) $A(t)$ of the background star depends on the projected distance (in the plane observed) between that star and the foreground system $u(t)$ during a passage in the following way (following e.g. the procedure in Bennett & Rhie, 1996):

$$A = \frac{u^2 + 2}{u\sqrt{u^2 + 4}} \tag{4}$$

$$\mathbf{u} = \frac{(t - t_0)}{t_\mathrm{E}}\mathbf{x} + u_0\mathbf{y} \tag{5}$$

where u_0 is the shortest projected distance during the passage, t_0 is the corresponding point of time and t_E is the so-called Einstein time scale, which depends (though not uniquely) on the mass. If the foreground mass is just a single object, then the magnification curve as a function of time will just vary smoothly with a maximum value at t_0. However, if the object has a satellite (e.g. a star with a planet) and the satellite at some point during the passage happens to be positioned in such a way that a so-called caustic (a situation in which the foreground object/s are placed such that u is zero, resulting in a large (theoretically infinite) magnification) is produced, there will be a sharper peak in the curve at the time where the caustic took place (see figure 6).

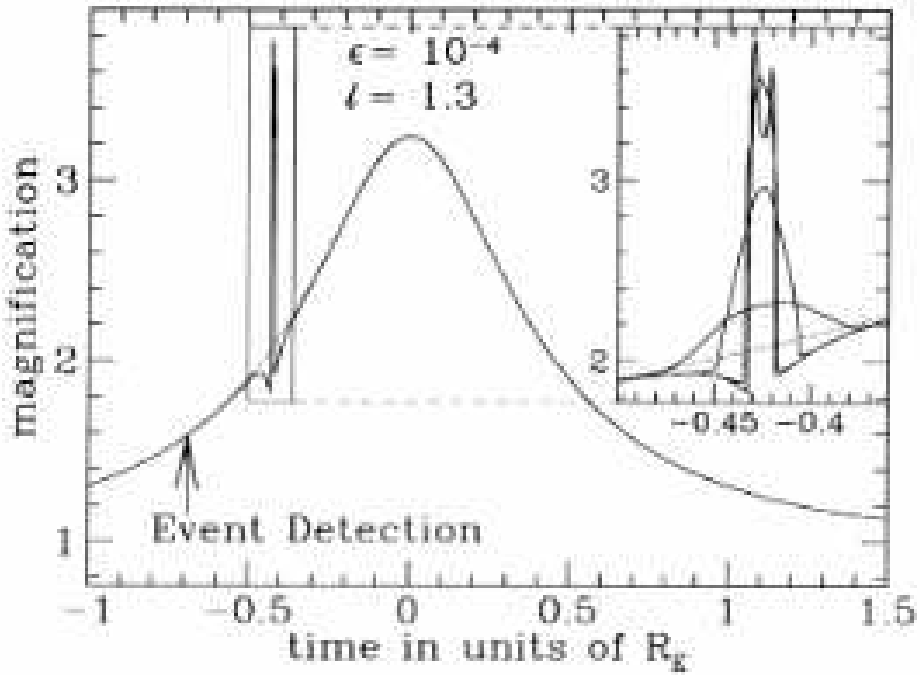

Figure 6: Example of a caustic in an otherwise smooth curve (from Bennet & Rhie 1996).

Several such caustics have indeed been observed during lensing events, indicating that the lensing star in the various cases could be harboring planets. However, the results are generally ambigous, since the Einstein time scale does not uniquely depend on mass, but also on proper motion and the relative source-lens parallax. An event with a short Einstein time scale could be a planet with low proper motion, but also e.g. a brown dwarf with high proper motion. It is therefore most often not possible to conclude the existence of a planet with this method – only one certain planet (see Bond et al. 2004) has been deduced among the many candidate lensing events observed so far. A solution to this problem has been suggested by Han et al. (2004) who note that aside from the Einstein time, two other observables of the lensing events (the Einstein ring radius and the Einstein radius projected onto the plane of the observer) can be measured using two telescopes

with a sufficiently large separation (such as one space-based and one ground-based). This gives enough information to solve for a unique mass behind a certain lensing event.

Using the above modification of the method, one can also detect free-floating planets – an application which no other known method can provide. Free-floating planets are interesting from the perspective of planetary formation, since some models predict that many planets should be expelled from the system in which they originate through gravitational interaction with other planets (see more about this in the "Extrasolar planet properties" section). Thorough gravitational lensing studies could provide observational constraints on this.

Gravitational lensing is one of the few techniques which can be used to detect planets on the order of Earth's size, although detection probability decreases with decreasing size of the planet. It also works over larger distances than the other methods. The big disadvantage, however, is that lensing events are extremely rare. To maximize the probability of detecting an event, one should look towards a direction in which as large an amount of potential background sources as possible are concentrated, i.e. towards the galactic centre. Even so, lensing events are few in amount and can't be predicted (especially not for free-floating planets), which on the whole makes gravitational microlensing a promising but unflexible method.

3.5 Direct Imaging

Perhaps the most informative and certainly the most intuitive method to detect extrasolar planets is to directly study the light that they emit. This is how we gain most information of stars, pulsars, galaxies and planets in our own solar system. For extrasolar planets, though, this is easier said than done. These objects are not only poorly luminous in themselves, they are also positioned close to a harboring star which outshines them by about a factor 10^9 in the visible range. One can decrease this monstrous difference somewhat by instead observing in the infrared frequency range, where the star is "only" a factor of about 10^6 times brighter. Finding a planet under these conditions is still completely impossible by any currently existing telescope, ground- or space based, regardless of resolution.

Several suggestions for dealing with this problem have been put forth, such as using coronographic masks to mechanically block out the light from the star (such as is traditionally done when studying the corona of our sun, thereby the name), but the most promising idea is to search for extrasolar planets using an interferometric set of telescopes. The basic principle for a two-element interferometer in such an application will be described below.

The two elements are directed straight towards a hypothetically planet-harboring star with a baseline of length D between them (see figure 7), observing at a wavelength λ. For a certain angle θ relative to this direction of the infalling light, the difference in path length that the light has to travel in order to reach the different telescopes is $D \sin \theta$. If we add the signal from the two telescopes together, we get positive interference when the path difference is equal to a whole amount of wavelengths, and negative interference for a half amount. That is, by combining the information from the two telescopes in a direct additive way we get interference fringes with intensity maxima for $D \sin \theta = n\lambda$ and minima for $D \sin \theta = (n/2 + 1/2)\lambda$ where $n = 0, 1, 2, ...$

If we in addition to this artificially add half a wavelength to the signal received by one of the telescopes such that its phase is shifted by π radians, we will also shift the interference pattern such that the angles resulting in minima without a phase shift instead

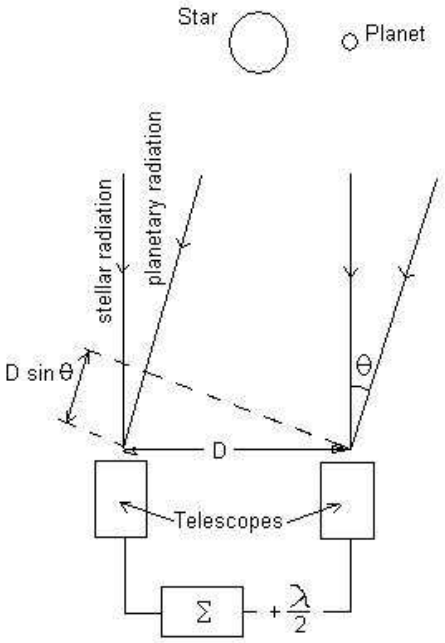

Figure 7: Principle of a two element nulling interferometer.

become maxima when a phase is added, and vice versa. With this setup, therefore, an intensity minimum will necessarily occur at an angle of zero – that is, corresponding to the placement of the star. In other words, we have effectively nulled out most of the light from the star while retaining the possibility to detect light near the interference maxima at certain distances from the star (which ones depends on the baseline length). This technique is called nulling interferometry, and would be hugely beneficial for this application.

Another advantage of interferometric telescopes (and indeed, the main motivation behind interferometric projects in general) is the improvement in resolution – a set of such telescopes are equivalent (in terms of resolution) to one single telescope whose lens is as big as the longest baseline length in the interferometry set.

The technique described above will be used in Darwin, which is a project planned by ESA (see Fridlund, 2000). Darwin will be space based to avoid atmospheric effects and observe in the infrared for reasons mentioned. While the simple principle of nulling interferometry described contains two elements and therefore only can observe in a plane parallell to the baseline (generally corresponding to a line in the plane of the observed system), Darwin will consist of about six telescopes providing a two-dimensional view of any system. It should be able to detect planets all the way down to Earth scale. Darwin is projected for launch somewhere around 2014. A similar project planned by NASA called the Terrestrial Planet Finder (TPF) may use nulling interferometry, or possibly the coronagraphic mask method mentioned above if such should turn out to be technologically feasible (see e.g. Schneider 2004b).

Direct imaging of a planet provides huge benefits. The size of the planet is given directly, and through continous studies all orbital parametres can be determined. In addition, as is the case for transits, spectroscopical analysis can be performed, giving information about the planetary atmosphere. The main disadvantage of direct imaging is its sensitivity to distances.

3.6 Pulsar Timing

While the first extrasolar planet around a sun-like star was detected in 1995 using radial velocity, the actual first discovery of an extrasolar planet was made three years earlier using the method of pulsar timing (see Wolszczan & Frail, 1992).

A pulsar is a kind of supernova remnant. It is essentially a rapidly rotating neutron star with a strong magnetic field with its pole axis tilted to the rotation axis. The pulsar emits a strong, well collimated beam in the radio range along the magnetic field axis in both directions. For a distant radio observer close to the extension of this line, the pulsar will appear as a strong, short pulse every time the pulsar completes one rotation (which is quite often – the rotational period of pulsars range from the order of seconds to the order of milliseconds).

Since the pulsar rotates at a constant pace, the pulses will be received at a constant pace if the pulsar is unaccelerated. If, however, the pulses are received at a rate which decreases and increases in a periodical manner, we can conclude that a satellite revolves around the pulsar and determine some of its characteristics. This is the idea behind planet finding through pulsar timing.

The amplitude of the variations of pulse arrival time is given (assuming a typical pulsar mass of $1.35 M_{\text{sun}}$) by:

$$\tau = 1.2\left(\frac{M_{\text{p}} \sin i}{M_{\text{Earth}}}\right)\left(\frac{P}{1\text{yr}}\right)^{2/3} \text{ ms} \tag{6}$$

which means that if the pulsar is spinning fast enough (on the order of 1 ms), Earth-mass planets are well within reach for detection with this method. Since the periodicity can be observed, the remaining unknowns are the mass and inclination of the planet. Thus in general, only a lower mass limit can be imposed to the planet. This method is, as can be seen, in many ways similar to radial velocity, but with a vastly higher precision. The downside is of course the very limited usability. Pulsars are few in amount and in addition, pulsar planets most likely have to form by a separate process from the ordinary planetary formation scenario. If the planets had formed back when the pulsar was a main sequance star, they would have had to survive a supernova explosion, which is not very likely. For this reason, the amount of useful information pulsar timing could give us about general planetary formation is highly limited (although pulsar planet formation can of course be seen as an interesting area in itself).

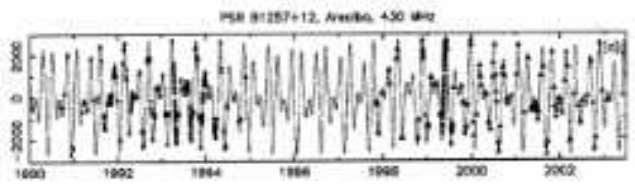

Figure 8: Timing of PSR 1257+12. Notice the excellent precision (from Konacki & Wolszczan 2003).

PSR 1257+12, around which the first planet was detected, is now known to harbor at least four planets, all detected with the Arecibo telescope. The three inner planets form a sub-system similar to the inner planets of our own solar system in terms of orbits and masses. Although in general, only the lower mass of planets is determinable through pulsar timing, the precision of the method is such that the gravitational perturbation exhibited

by the second and third inner planets on each other has been measurable, which yielded unique masses and therefore also inclinations of these planets (see Konacki & Wolszczan, 2003). The masses were found to be $4.3M_{\text{Earth}}$ and $3.9M_{\text{Earth}}$ at semi-major axes 1.3106 AU and 1.4134 AU, respectively. They also both have very low eccentricity and are in other words similar to Earth in this sense.

3.7 Other Methods?

Given so far are all the methods that have missions connected to them with a serious projected possibility to discover extrasolar planets. There may also be other methods, seriously considered or not, which could provide useful information about extrasolar planetary systems and their formation. One possible such idea is to study radiation emitted from planetary collisions. In Zhang & Sigurdsson (2004), this method has been investigated theoretically, and it was found that a collision between a terrestrial-type planet and a giant-type planet would produce a specific kind of light signature (a short X-ray flash followed by a drawn out IR afterglow) which should be detectable from Earth under some circumstances. Such collisions would be most frequent in newly formed planetary systems, and if they are common (and detectable as theory suggests), observations of them could for instance provide information about creation of moons (which have been suggested as in many cases having been formed through collisions).

4 Properties of Extrasolar Planets

Having gone through the theories about planetary formation, as well as the techniques used for detecting planets in systems separate from our own, it is now time for us to see in more detail what these observations have shown, and how it affects the theories. It turns out that a vast majority of the extrasolar planets detected so far can be placed in one of two categories: planets with a high mass and very small, near-circular orbits (hot Jupiters) and planets with high mass and moderate-size orbits with very high eccentricities. These categories and some possible mechanisms behind them will be discussed below, and thereafter some properties of planetary systems as seen from their harboring star/s will be described.

4.1 Hot Jupiters

Planets of this category have semi-major axes of around 0.05 AU (much smaller orbits than Mercury which has a semi-major axis of about 0.4 AU) and masses on the order of Jupiter's. Technically, of course, only a lower limit for the mass is known, but statistically (if we assume that the inclinations are randomly distributed), the unique mass should remain at this order of magnitude on average. In addition, transit measurements (see e.g. Konacki et al. 2003 and the "Transits" section of this paper) have shown the concept of hot Jupiters to be relevant in at least two cases. In addition to the mass and orbit size, the orbits have in common a low eccentricity (from about 0.1 and downwards).

That planets of this kind have been found with radial velocity searches is not strange *per se* (because as we have seen, radial velocity is biased towards both high masses and small orbits), but it *is* strange that they exist in the first place to such an extent. The problem is, that the accretion disc is thinner close to the parent star than further out, and for this reason, there should not be enough material close in to form giant planets

in general. This problem remains independently on which mechanism is assumed behind giant planet formation (core accretion or disc instabilities).

The only way to save either of the two formation scenarios is to postulate that hot Jupiters form further out in the disc (around the current distances of the giant planets in our solar system) and then travel inwards until they reach the distances from the star as observed today. Suggested mechanisms behind such a process is viscous action from the disc before it evaporates, causing the giant planet to slow down and travel to smaller orbits gradually. Another idea involves gravitational interactions between the giant planet and other planets or planetesimals throwing the giant into an elliptical orbit which is then circularized (at a closer distance from the star than the original orbit) through tidal interactions with its star.

The introduction of migration as a solution to the hot Jupiter problem in turn introduces another problem; the rate of inwards motion should increase with smaller orbit, which means that whatever mechanism causes the planet to travel inwards towards the star should consequently cause it to plunge into the star and be destroyed. Instead, the frequency of hot Jupiters indicate that the migration suddenly stops before this takes place. One must thus introduce yet another mechanism to deal with the halting of inwards migration at small orbital radii. Examples of candidate processes are clearing of the inner disc through interaction with the star's magnetic field (removing viscous interactions in this region), and mass loss from the planet to the star through the Roche lobe, leading to a counteracting outwards angular momentum distribution (as has been mentioned in the "Transits" section, HD 209458 has been observed to lose mass to its harboring star through the Roche lobe).

4.2 Highly Eccentric Orbits

Planets forming in a disc should in general have very circular orbits, such as is the case in our own solar system. However, a great majority of all extrasolar planets found with a semi-major axis above about 0.1 AU (and below the semi-major axis of the largest extrasolar orbit found to date at 4.8 AU) have greatly eccentric orbits (not rarely larger that 0.3, record being HD 80606 b with an eccentricity greater than 0.9!). In order to get the standard giant formation models to apply, we must introduce a mechanism with the ability to excite eccentric orbits from initially circular ones.

A couple of suggestions for such can be found in e.g. Tremaine & Zakamska (2003). One such takes into consideration resonant interaction between the planet and the accretion disc (that is, orbital properties of the planet resonate with orbital properties of the disc as a whole) which could excite low eccentricity orbits (but also damp high eccentricity orbits). Another possible eccentricity-producing mechanism is the gravitational effect of a close interaction with another planet (similar to one of the hot Jupiter explanations mentioned, but without tidal circularization effects). Yet another possibility is continous influence from a massive nearby body that is not co-planar with the planet-star system (for instance a binary component to the harboring star). No mechanism alone seems to be able to quantitatively explain all the eccentric orbits and their corresponding system for all observed planets (for instance, a non-co-planar binary component could explain the eccentricity for some cases, but certainly not all, unless it hides its presence really well such as through a non-inclined orbit in an awful lot of systems). Rather, it is probably a question of different processes at work in different systems.

A separate hypothesis is that rather than forming in the way of the gas giants in our system (wether that was core accretion, disc instabilities or anything else), the extrasolar

planets observed so far were created along with the star in the initial collapse of the system. Theoretically, this would cause planets formed this way to have either very short orbits with small eccentricities or medium size orbits with high eccentricities. This would elegantly explain all orbits of extrasolar planets observed so far as well as eliminate the migration problem. However, there is a major setback to this theory: observed extrasolar planets seem to have an average mass slightly lower than Jupiter's, whereas fragmentation during the stellar collapse craves a mass of at least a few Jupiter masses.

4.3 Metallicity

Statistical analysis indicates that stars harboring extrasolar planets on average show a higher amount of metals in their spectra than stars with no planets found. This is generally interpreted as that high metallicity content in the disc favours the formation of giant planets, although there have been suggestions that the connection goes the other way around – that presence of planets favour detection of high metallicity. The thought behind the latter suggestion is that stars with a lot of planets will be more prone to collide with their planets which could increase the concentration of metals in the envelope of the star, which is what shows in spectroscopical surveys (rather than the metal content of the star overall). However, the first suggestion seems more reasonable on the whole and is somewhat supported by transit observations of a dense, metal-poor cluster (see Gilliland et al. 2000 and the "Transits" section of this paper). Several planets were expected in this survey assuming the same fraction of planet-harboring stars as in the solar neighbourhood (as found by radial velocity searches), but not a single one was found.

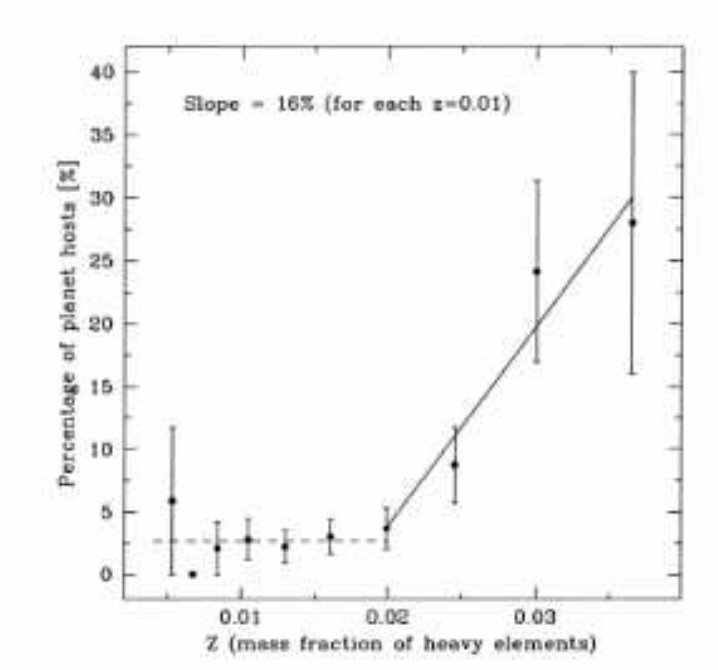

Figure 9: A diagram of planet hosts against metallicity. There is a clear correlation for most part, but possibly also a metal-independent component (from Santos, Israelian & Mayor 2004).

A connection between high metallicity and high frequency of massive planets should speak in favour of the core accretion model, because high metallicity in the accretion disc

means availability of more dust grains to form the cores which later become gas giants. An interesting alternative is presented in Santos et al. (2004), where it is noted that aside from an increasing amount of planet-harboring stars with increasing metallicity, a metallicity-independent component can be read out from the data (see figure 9). This could be interpreted as two separate populations of stars – one formed by core accretion and one by disc instabilities. There are two weaknesses in such a reasoning, though: first, the metal-poor planet-harboring stars are few in amount and thus a statistically uncertain group. Also, aside from metallicity, there are no statistical differences between the two hypothetical populations of stars, which could perhaps be expected if they formed by entirely different processes.

It is important to remember that the observed correlation between metallicity and planet presence applies uniquely to the kinds of planets found – hot Jupiters and giant planets at highly eccentric orbits. We can draw no such conclusion for terrestrial planets, or necessarily even for the planets equivalent to the Jovian planets of our solar system. For reference, the sun is an average metallicity star in its neighbourhood and thus below the average metallicity of known planet-harboring stars, and still it has nine planets of which four are giants.

5 Terrestrial Planets

So far, the focus of this paper has been on giant planets. This has its simple explanation in that no smaller extrasolar planets have been detected, nor can be detected by presently existing instruments (not without an extraordinary amount of luck, anyway). Properties and frequency of extrasolar terrestrial planets is therefore still an entirely theoretical and by necessity somewhat speculative field of research. Nonetheless, it is a very interesting subject, and partly since this paper is about extrasolar planets as a whole and partly because a discussion of places which could harbor life might be of particular interest to the reader, there will follow a discussion on Earth-like bodies (in terms of size and distance from the star) in relation to known extrasolar systems.

It is often stated that the extrasolar systems found to date are very badly compatible with Earth-like planets. This is true for most part. For instance, a planet such as HD 2039 b which has a mass of $4.85 M_{\mathrm{J}}$, semi-major axis of 2.19 AU and eccentricity of 0.68 would efficiently wipe out any terrestrial planets in the habitable zone (the distance range from the star in which a planet can maintain fluid water – basically stretching from Venus to Mars for a sun-like star) that might once have been there through gravitational ejection or collisions. However, all detected systems are not this hopeless. One can sort out three different kinds of detected systems that could harbor Earth-like bodies in stable orbits, which will be described below.

Hot Jupiters do not significantly affect planets in the habitable zone in their present state. However, the possibility of existance of terrestrial planets in the habitable zone is often excluded because by standard models, hot Jupiters were formed further out in the disc and then migrated inwards, and it is commonly believed that terrestrial planets could not survive such a migration but rather be thrown out of the system when the giant planet comes close to their orbit. This is not necessarily the case. Simulations by Mandell & Sigurdsson (2004) show that a small fraction of the affected planets end up in an orbit close to their original one after the migration is finished. The fraction increases with increasing speed of the migration. If the migration lasts for 2 Myr, about 10 % of the terrestrial planets stay in the system and for 0.5 Myr, about 40 % remains (though

not necessarily in the habitable zone).

There have also been findings of giant planets which are far out enough and non-eccentric enough to have terrestrial planets on stable orbits inside of them. Simulations have been done for five of the most promising systems in this aspect (Asghari et al. 2004), and it was found that terrestrial planets have good chances of staying in stable orbits within the habitable zone in four of them (in particular Gl 777 A, wherein the giant planet has the largest orbit of all detected extrasolar planets).

Finally, there is a possibility for earth-like objects to reside in the habitable zone not as planets, but as moons to a larger planet. This giant must of course itself have the appropriate semi-major axis and eccentricity to stay within the habitable zone for this to occur. A possible detected giant candidate suggested in literature (Barnes & O'Brien 2002) is HD 28185 b which has a semi-major axis of 1 AU and eccentricity of only 0.06. This kind of setup is the only one (of the three mentioned) in which a habitable body could actually be detected (with quite a bit of luck) using present techniques (see the "Transits" section).

It is important to note that the above discussion deals exclusively with planetary systems of the kinds that have been detected. The average planetary system may look completely different, leading to a better or perhaps even worse situation for the frequency of terrestrial planets. See the beginning of the "Discussion" section for a further development of this point.

6 Discussion

There seems to be a somewhat common conception that the systems detected so far constitute the norm of planetary systems in general, and that the majority of all solar systems thus contain hot Jupiters or giants in highly eccentric orbits. This is completely unfounded. It is difficult to find estimates in literature of the percentage of solar-type stars that have these kinds of planets, or even the percentage of stars investigated (a number of 5 % is given in Boss (2002b), but the context is somewhat unclear). I have therefore made an estimate based on results from the ELODIE project. In this project, 330 stars have been monitored since 1996 (see Naef et al. 2004). Counting from a table of detections in Udry (2004), it is found that 28 planets have been detected (counting both primary, secondary and independent simultaneous detections). Given the fine precission and long observational time of ELODIE, it is assumed that all giant planets within a semi-major axis of about 4 AU and inclination of more than 45 degrees (give or take on each side depending on the actual mass of the object) have been detected. The estimated fraction of hot Jupiters and giants with high eccentricities to total amount of solar-type stars is then given as 11.3 %.

The point of this procedure is not to give an exact amount of such a fraction, or even necessarily a good estimate. The actual number could probably be both much smaller and much larger depending on how the assumptions relate to reality. Rather, the whole point is to underline the fact that noone really knows how a typical planetary system looks yet. It could consist of only giant planets on eccentric orbits, or it could look similar to our own system – the point is that there is not a sufficient basis to draw conclusions either way from the data currently available.

The issue of core accretion versus disc instabilities remains undisclosed. While core accretion currently seems to make a somewhat stronger case, this could turn around as the knowledge of extrasolar systems and accretion discs increases.

Extrasolar planets is a young field of research, and technologically demanding. As time goes on, the presently existing radial velocity searches will keep detecting giant planets further out from stars if they are common, but their capacity ends at Jupiter analogues, with further technological development aiming at Saturn analogues. Present astrometry has limited capacity, but planned projects seem to be able to make the method into a good complement to radial velocity with an ability to detect giant planets on large orbits, maybe down to Uranus equivalents, and ability to determine a unique mass in combination with radial velocity data. Transits and gravitational microlensing will probably provide high quality data for a small amount of planets in the future, perhaps with increasing frequency as technology improves. Eventually though, it will probably be space-based interferometers such as Darwin that give the first somewhat detailed view of an extrasolar system with its projected ability to detect planets down to earth-size at a wide range of orbit sizes through direct imaging. What such surveys will show, nobody knows. An exciting 20 years to come is to be expected within the field of extrasolar planetary science.

Acknowledgements

The author wishes to thank Alessandro Romeo for his support. Also, thanks go out to Eva Karlsson whose proceedings from a previous session of the course provided some valuable references that helped me get started.

References

Asghari N. et al., 2004, submitted to A&A
Barnes J., O'Brien D., 2002, ApJ 575, 1087
Benedict G. et al., 2002, ApJL 581, L115
Bennett D., Rhie, S. H., 1996, ApJ 472, 660
Bond I. A. et al., 2004, ApJL 606, L155
Boss A., 2002a, ApJ 576, 462
Boss A., 2002b, Earth Planet. Sci. Lett. 202, 513
Brown T. et al., 2001, ApJ 552, 699
Charbonneau D. et al., 2000, ApJL 529, L45
Gilliland R. et al., 2000, ApJL 545, L47
Han C. et al., 2004, ApJ 604, 372
Konacki M et al., 2003, Nature 421, L507
Konacki M., Wolszczan A., 2003, ApJL 591, L147
Mandell A., Sigurdsson S., 2003, ApJL 599, L111
Mayer L. et al., 2003, ASP Conf. Series.
Mayor M. et al., 2004, A&A 415, 391
Mayor M., Queloz D., 1995, Nature 378, 355
Naef D. et al., 2004, A&A 414, 351
Perryman M. A. C., 2000, Rep. Prog. Phys 63, 1209
Rice W. K. M., Armitage P., 2003, ApJL 598, L55

Santos N. C., Israelian G., Mayor M., 2004, A&A 415, 1153

Schneider J., 2004a, The Extrasolar Planets Encyclopaedia:
`http://www.obspm.fr/encycl/encycl.html`

Schneider J., 2004b, ESA SP 539, 205

Tremaine S., Zakamska N., 2003, Invited review at "Research for other worlds", College Park, MD

Udry S., 2004, The Geneva Extrasolar Planet Search Programmes:
`http://obswww.unige.ch/~udry/planet/planet.html`

Vidal-Madjar A. et al., 2003, Nature 422, L143

Vidal-Madjar A. et al., 2004, ApJL 604, L69

Wolszczan A., Frail D. A., 1992, Nature 355, 155

Zeilik M., Gregory S., 1998, Introductory Astronomy and Astrophysics. Saunders College Publishing

Zhang B., Sigurdsson S., 2004, ApJL 596, L95

Life in the Universe

Mats Johansson

Göteborg University
SE-41296 Göteborg, Sweden
(hallonsnok@hotmail.com)

*

Abstract

This paper discusses which properties are necessary for life. It starts with an attempt to give a definition of life and what life needs. It continues with a discussion on the importance of organic molecules in DNA and RNA. I will discuss which regions in the universe that are interesting and I will then suggest different properties (such as mass of stars, the habital zone and more) that will be important. Then a look at the recent findings of life in extreme environments, gives a different approach to the habital zone. A brief look into the interstellar chemistry reveals important molecules in the interstellar medium and comets. The paper also discusses how these molecules could come to earth and how the first cell evolves. Finally there is a discussion about earth and if it is unique or not and if the evolution from single celled life to multiple celled and intelligent life is unique for earth.

1 Introduction

Life at other places in the universe have fascinated people for a long time. Fantasies about life on Venus, Mars and even the moon have been on peoples minds. When early astronomers looked at the surface of Mars, they saw something that they explained as water channels, they said that these channels were so precise that only intelligent beings could have made them. Later on when pictures on the surface of Mars were taken, people claimed that they could see a huge face. This could of course not be a coincidence, there simply must exist or at least have existed life on Mars to construct this face. Today we know that the so called water channels were not handmade but rather made from natural processes, however it is still possible that water were present in them. The so called face has turned out to be a mountain and for different reasons appeared as a face on the pictures. We still do not have any sure evidence for life, besides life on earth.

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

2 What Is Life and What Does It Need?

The first thing we have to do, is to establish some sort of definitions. It is hard to give a simple definition of life, because it is usually easy to give a counterexample to the definition. However, we still need to have a few rules.

2.1 Metabolism

The cells need energy to produce aminoacids and other important molecules. The metabolism supplies the cell with energy for the different processes.

2.2 Reproduction

It is important that living organisms can produce new individuals. These should have similar properties compared to the original organisms.

2.3 Complexity

A certain degree of complexity is needed, although there is not a well defined limit of complexity. An example would be a crystal that copies itself. This would not be complex enough.

2.4 Evolution Due to Mutation or Adaption

The ability to evolve may not be necessary to a single organism, however it will be more important for an entire species. The mutation appears when the reproduction is not perfect, new abilities will appear. The adaption will occur over a longer time period and new abilities will slowly appear, perhaps due to changes in the environment.

2.5 The Ability to Store and Transfer Information

The information in the cells are processed with DNA (Deoxyribo Nucleic Acid) and RNA (Ribo Nucleic Acid). DNA stores the information and it is built up by the four bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). A and T, G and C match up to be base pairs. Each base is connected to a deoxyribose (C5H10O4) and between two of these, there is a connecting phosphorus. The RNA uses the information to produce proteins and other molecules. RNA also uses four bases but instead of Thymine, RNA uses Uracil (U) and instead of deoxyribose, RNA uses Ribose (C5H10O5). In order to produce the bases in DNA and RNA, there need to be proteins. To synthesize proteins there has to be aminoacids. Aminoacids are built up by a main carbon atom with four different attachments. The first attachment is a hydrogen atom, the second is an aminogroup (NH3), the third is a carboxylgroup (COOH) and the last attachment is one of twenty different R-groups. The R-groups decide which aminoacid it is, the aminoacid alanine has the R-group CH3. Aminoacids are organic molecules, other examples of organic molecules are methane (CH4) and ethane (C2H6), these are simple organic molecules.

2.6 Boundary

To protect everything some sort of boundary is needed. Cells have a membrane consisting of hydrophobic molecules, these molecules create a protected area.

2.7 Water

Water is a good environment (good solvent) in which the different parts and molecules can mix and create the necessary ingredients for life.

2.8 Energy

The metabolism can use either organic molecules (such as sugar) as its energy source or it could use nonorganic molecules (such as S, Fe or H2) as energy source.

2.9 Safe Environment

In order for life to evolve and exist, the environment must not be to rough. The temperature must not be to low or to high. There are also restrictions on the radiation, the UV-radiation must not be to severe because then long molecules such as the aminoacids would brake apart. It is also important that the frequency of impacts from meteorites, comets and asteroids is not too high, if it is too high it would definitely destroy all opportunities for life to evolve. High volcanic activity would also have the same effect.

3 Regions Where Life Can Exist

It is a good idea to try to narrow down the places where life can evolve, that means to try to find properties that are necessary for life or are hazardous for life.

3.1 Right Kind of Galaxy

The question of which galaxies that are most suitable to harbor life depends mostly on the fraction of metallicity. A low fraction of metals would imply that terrestrial planets (or satellites) could not be able to form. The earliest galaxies have a very low metalicity, they would also have dangerous quasar like activity and several supernova explosions. This makes the earliest galaxies unlikely to harbor any form of life. Small galaxies also have a low fraction of metallicity and they are therefore not suitable for life to form. Globular clusters, that contain up to a million stars and is spread around in the halos of galaxies, have been shown to be metal poor and containing mostly old stars. However searches for planets have been made and it was expected that several planets would be found, but they did not find any planets. One possibility why no planets were found could be that the globular clusters are to compact, so that if planets have existed, they have been disturbed and pushed away from their parent star. Another possibility is that the fraction of heavy elements are to low.

3.2 Position in the Galaxy

In the centre of galaxies energetic processes occur that are hazardous to life. It is also possible that stars collide close to the centre. One can safely say that the centre of a galaxy is too violent to harbor life. Further out at the edges and in the halos of galaxies, most stars are too metal poor and it is unlikely that there exist enough of the right materials to form terrestrial planets (or satellites).

3.3 Right Kind of Star

The most important property of the star is the mass. Since the mass affects the lifetime of the star, it is therefore important that the mass is low enough. The sun has a lifetime of 10 billion years, a star with 15 times the solar mass has a lifetime of 10 million years and a star with 0.25 times the solar mass has a lifetime of 70 billion years. Altough the lifetime has to be long enough, it is also important that the star can supply the terrestrial planets with enough energy. The mass should therefore not be too low or too high.

3.4 Stable Planetary Orbits

If the planetary orbits are unstable, it is likely that one or more planets will be forced away from its parent star or plunge into the star. It is therefore unlikely that such planetary systems are suitible for life. If the orbit of a planet is to eccentric, the environment on the planet would change so much that it would be likely that it is to extreme for any life to survive or even to evolve. If there are giant planets in the system then it is more likely that the entire system has stable orbits with reasonably low eccentricity.

3.5 The Habital Zone (HZ)

The habital zone is defined as the space around a star where liguid water can exist due to heating from the star and the temperature on the planet allows an atmosphere. The atmosphere is needed to protect the surface from stellar winds, high-energy particles, dangerous UV-radiation and more. Life is believed to require a liquid environment, water is a very good liquid for this purpose. The reason why water is good is that it is a very good solvent. Also important is that cell membranes that protect living organisms from their surroundings are destroyed at 150 degrees centigrade and if the inner temperature reaches below -10 degrees centigrade, the cell goes to coma. These limits are obviously close to the limits of liquid water. The habital zone for the sun is between HZ = 0,7 AU and HZ = 1,5 AU. For a general star the habital zone can be calculated from:

$$HZ(star) = HZ(sun) * [L(star)/L(sun)]^{1/2}$$

Where HZ is the distance to the star and the sun, L is the luminosity of the star and the sun. This habital zone could be extended if a planet has a thick enough atmosphere, then a planet that should be too far away could still have liquid water.

4 Life in Extreme Environment

Recent explorations have discovered that life can survive in places that previously was thought to be too hot, too cold or too dry. These discoveries have opened new possibilities to find life elsewhere in the solar system and the universe.

4.1 Deep Oceans near Vulcanic Sources

When searching deep in oceans where the water is heated by volcanic activity (up to 380 degrees centigrade), bacteria have been found. The hot water contains chemicals and minerals, where the hot water mix with cooler water the bacteria thrives. The energy sources that these bacteria use are H2, H2S, CO and Mn+. The bacteria do not use

photosynthesis nor do they use organic energy source, they are therefore referred to as chemoautotrophs. These bacteria that live in hot water are also called hyperthermophiles, they live in water that is hotter than 80 degrees centigrade. The reason why the water does not boil at these high temperatures is due to high pressure. Another type of hyperthermophiles exists in hot springs at the surface.

4.2 Arctic Conditions

Bacteria have found a way to live in dry, cold valleys in Antarctica. The temperature does only rise above the freezing point of water for a few days a year and there is low abundance of water. The way in which the bacteria survive is to be inside rocks. The temperature may rise above the freezing point inside the stone even if the air temperature is below, this is possible because a rock is good at absorbing the sunlight. A rock also contains small spores in which water, when available, can diffuse into. Other parts of the rock are made of transparent mineral grains, this opens a possibility for sunlight to penetrate short distances of the rock. These bacteria are called lithoautotrophs, litho because they live inside rocks and autotrophs because they use CO2 from the atmosphere.

4.3 Deep Inside Mountains

Bacteria have been found in vulcanic rocks deep below the surface, these bacteria use iron as their energy source. They are usually found in acid conditions and they are using hydrogen to dissolve nonorganic carbon into reduced forms and even into organic molecules such as methane. The hydrogen that is needed for this comes from a chemical interaction between water and iron. These bacteria do not rely on a previously existing organic carbon source, instead they are able to produce it. This is very interesting because neither sunlight nor organic molecules are necessary.

4.4 How Does This Affect the Habital Zone?

Water is still an important element, and it is also important that it can exist in its liquid form. However if it is possible to have another source of energy than the star to heat the water, then it is possible to have liquid water outside the previously mentioned habital zone. Two examples of such energies are volcanic activity and gravitational pull from a planet on an orbiting satellite. If the water is beneath the surface of the planet (or satellite), then the effect of a protecting atmosphere could get less important. An interesting example can be found in the solar system, the satellite Europa that orbits around Jupiter is clearly outside the habital zone. Europa seems to have a surface of ice, due to a big gravitational pull from Jupiter the surface has huge cracks on the surface, although the cracks are frozen it could still mean that there is enough energy to get liquid water beneath the frozen surface. If there is a huge ocean on Europa and there are hot sources such as vulcanic activity on the ocean floor, then it is possible that life has evolved in the ocean on Europa. Europa is therefore a great example that shows that the habital zone can be extended a lot.

5 Where Did (Does) Life Start?

The most important molecules are made up from carbon, hydrogen, nitrogen, oxygen, phosphorus and sulphur, then we have water that acts as the solvent. However molecules

can also be made elsewhere in the universe. To create complex molecules there has to be enough time for the different processes to occur. This is solved by using the surface of grains, which is created in the rapidly cooling halos of ageing and expanding stars. These processes, however, do not explain the large abundace of the most complex molecules.

5.1 Important Molecules in the Interstellar Medium and Comets

Observations of the interstellar medium and comets have shown a large abundance of both nonorganic and organic molecules. A few examples are CH4, C2O, H2O, CH3OH and many more, molecules as large as 13 atoms have been found. The long carbon chains HCN, HC3N, HC5N up to HC11N are very interesting. The shortest of them (HCN) can be used to create adenine (H5C5N5). Recent simulations have shown that a significant amount of adenine could be produced in the collapse of molecular cloudes, this is during a time period of 1-10 million years. Since adenine is one of the bases in DNA and RNA this is very interesting. Another interesting molecule is the ring molecule C2H4O, this molecule is very similar to another ringmolecule C4H4O and this molecule is an important part of both deoxyribose and ribose. This shows that it is possible that important molecules are created and could exist in the interstellar medium and comets. Water is found to be abundant in the interstellar medium and especially in comets.

5.2 Young Earth

During the early period of earth's history, several impacts occured. Since comets contain huge amounts of water, several organic molecules and other important building blocks for life, it is likely that most of the water on earth came from comets and also some of the important CO2 is likely to have come from comets. Comets are likely to be common in any planetary system and it is therefore likely that comets would crash into planets in other systems as well. This would then imply that other planets would have access to big amounts of water, organic molecules and other important molecules.

5.3 Beginning of Life on Earth

Assuming that life evolved on earth and was not planted here, where could it have begun? Several places are suggested, ocean floor near hot sources, waterfilled cracks deep below the surface and water near or on the surface, these are some ideas. Hot water from deep beneath the surface will mix on its way up with sulphur, hydrogen gas, ammonia, methane, carbon dioxide, metals and salts. As shown earlier there are bacteria that can produce organic molecules from CO2, so there is no need for organic molecules to exist prior to these bacteria. We also know that comets have brought organic molecules to earth, so bacteria that use this could also evolve at this time. Since there was no oxygen in the early atmosphere on earth, there was not any ozone to protect the surface from dangerous UV-radiation. There were also several impacts on the young earth, all of this would make the surface a dangerous place for life to evolve. There is no clear evidence on where life evolved or if it evolved with an origin of nonorganic or organic energy source. We do know that organic molecules such as aminoacids are necessary for DNA and RNA. In order for complex organic molecules to form in a cell, there need to be some sort of protection, a membrane. The membrane then protect the cell and all parts and properties it contains.

6 Is Earth Unique?

Previous discussions would imply that single celled life could be rather common throughout the universe. The question now is if multiple celled life and intelligent life is unique here on earth or if that also could be spread throughout the universe. To try and answer this it is necessary to see which the important properties and events are and if they are unique for earth.

6.1 Planet Size

The size of the planet is important because it has to be big enough to retain an atmosphere and an ocean. The planet must also have enough heat to have plate tectonics.

6.2 Big Neighbour Planet

It is important that there exists a big planet in the system. This planet should not be too close or too far away. On the right distance it will protect the inner planets from massive and frequent impacts of asteroids, meteorites and comets. A recent example of this is when the Shoemaker-Levy crashed into jupiter. Another important use of big planets in the system is that they help to stabilize the planetary orbits in the whole system.

6.3 Plate Tectonics

Plate tectonics is the movement of the continental plates. We know that several million years ago there was one huge continent (Pangea) and at present time there are several continents. These movements and some atmospherical processes cause the carbon cycle.

$CaSiO_3 + CO_2 <=> CaCO_3 + SiO_2$

This will work as a global thermostat, because the CO2 in the atmosphere will be kept relatively uniform. This is very important in allowing water to be liquid for more than 4 billion years. In order to have plate tectonics the planet must have a thin litosphere and continental sized plates.

6.4 Large Satellite

It is important to have the right tilt of the planet, so that the seasons do not get too severe. A large satellite will help to stabilize the tilt.

6.5 Evolution

Single celled life existed for at least 3 billion years before multiple celled life evolved. It is possible that there was some event that caused this evolution. it is also possible that single celled life needed that much time to find a functioning system for multiple celled life. If such a long time period is necessary to evolve multiple celled life, then it is very likely that there has to be similar properties as it was here on earth during this period. When multiple celled life finally evolved, there was an explosion of various species. Due to several catastrophic events, such as big asteroid impacts, surviving species evolved in many different ways to use the resources that had been used by species that had become

extinct. Intelligent Beings have only existed for a short time period in the evolution and precisely how intelligence evolved is not really known.

7 Conclusion

The conclusions we can make are that the properties that enables single celled life, such as chemical processes, properties of the planetary systems and the planets themselves, should be rather common. However the evolution from single celled life to multiple celled life and intelligent beings is probably much more rare. So if life exists anywhere besides on earth, it is most likely to be single celled.

Acknowledgements

I would like to thank Alessandro and all the students that did the workshop for their support. A special thanks to Joanna Wiberg for helping me with the final spelling check. I would also like to thank my family and friends for their support.

References

Berntsson, 2003, Hot Topics in Astrophysics 2002/2003, Chalmers University of Technology and Göteborgs University

Booth, Hjalmarsson, 1997, Radioastronomi och livets uppkomst, Swedish Science Press, Uppsala

Chakrabarti S, Chakrabarti S.K, 2000, astro-ph/0001079 accepted in Astronomy and Astrophysics Letters

Charnley, Huang, Kuan, Rodgers, 2001, Astro-ph/0104416, accepted by Advances in Space Research

Dyson, Williams, 1997, The physics of the interstellar medium, IOP Publishing Ltd, Bristol and Philadelphia

Jakosky, 1998, The search for life on other planets, CambridgeUniversity Press, Cambridge

NASA, http://www.astrochem.org/science.html

Organic Molecules in Comets

Niklas Vahlne

Chalmers University of Technology
SE-41296 Göteborg, Sweden
(f99niva@dd.chalmers.se)

*

Abstract

The possibility of organic compounds delivered to earth via comet impacts is an interesting field. This report is a brief introduction to the subject aimed more at being an overlook of the different aspects of this diverse field of research. Observations of a wide array of organic molecules in comets and in the interstellar medium have been made, and these are discussed together with possible implications. The evolution of comets is accounted for. The effects of impacts on organic compounds are given a brief survey. Panspermia theories, especially ones related to comets, are discussed. There is still a lack of information concerning the composition of comets, and a mission has been launched to actually land on a comet to take samples. A short description of the mission and possible aspects are discussed.

1 Introduction

The idea that building blocks for life has arrived from space in form of cometary impacts has gained in credibility in recent years. Earlier experiments conducted by Urey-Miller, have assumed a different atmosphere than new research suggests. The chemical environment on the early earth may not have been an ideal place for more complicated organic molecules to form.

Lately lots of these have been observed in the interstellar medium and in comets. The definition of organic molecules are just carbon based molecules, not all have prebiotic importance in themselves but the presence of organic molecules could also indicate the presence of prebiotic ones. A question is if these more complicated molecules (amino acids) can survive the impact on earth. This is now being studied using shock chemistry, and some early results indicate that a substantial fraction of the amino acids can actually survive. Comets have for quite some time been regarded as totally conserved remnants of the solar system from the time when it was formed, new research is being made to test this thesis. The result is that passing stars, supernovae explosions, cosmic rays etc can change the comets compositions. So the comets we observe today may not be exactly like those hitting the earth during the heavy bombardment era.

*Hot Topics in Astrophysics 2003/2004, Alessandro B. Romeo, Martin Nord & Markus Janson (Eds.), Chalmers University of Technology and Göteborg University, 2004.

The observations so far has not provided us with all the information needed to fully determine what kind of molecules are present in comets. Direct examination of the nucleus are needed and such a mission is already on its way.

It is still a problem for theories concerning the origin of life that life could start so early in earth's history. To solve this problem it is suggested that not only organic molecules arrived from space but life itself. These are the so called panspermia theories in which comets can play a part.

2 Formation of the Solar System

The current model for the formation of the solar system suggests that the sun and the planets were formed from the same nebula. If the collapsing solar nebula had some initial angular momentum this would give rise to a protosun surrounded by a disk of gas and dust, that wounded around the protosun. The surrounding disk then started to cool and some elements began to condense, condensed particles collided with each other and gradually built up bigger objects. Near the protosun it was still to hot for water to condense and the colliding particles were made of silicates, iron, nickel, calcium etc, this would eventually lead up to the earth like planets. Beyond 5 AU water began to condense and could be included in the growing objects or planetesimals and from about 30 AU also methane. Beyond the 5 AU limit the Jovian giants could form. Long before the planetesimals had formed into the present day planets, ignition of the protosun occurred and the sun was born. This created solar winds that blew away the particles that had not yet condensed. The water and other volatile molecules within 5 AU was blown out giving the Jovian giants lot of mass to accrete. This lead to the creation of the mini solar systems around them, and they could accrete large gas envelopes from their surroundings. Smaller planetesimals in this region that managed to stay clear of the giants still had their orbits greatly perturbed, and sent away on eccentric orbits. These are the comets. The above discussion implicates that the early earth lacked water and other volatiles. If this model is true the water on the earth must be a result from cometary impacts. Thus the early oceans must have had similar compositions as the comets, and the study of comets becomes a crucial key to the understanding of the origin of life.

2.1 Comet Basics

In 1950 Fred L.Whipple came up with the term "dirty snowballs" for describing comets. The term coming from that a comet mainly constitutes of different types of ices with dust grains embedded within. As the comets approach the inner solar system the ice starts to sublimate, releasing gas and dust that creates a coma surrounding the nucleus. An envelope of hydrogen is layered outside the coma. Typical dimensions for these are 10^{11} cm for the coma and 10^{12} cm for the envelope. The nucleus itself is no larger than 10^{6} cm across. The dust particles are being hit by radiation pressure and a dust tail is created. Ions in the coma interact with the charged particles in the solar wind and create a ion tail. The ion tail always point directly away from the sun whereas the dust tail is slightly bent, due to the decrease in orbital speed for particles when the distance to the sun increases. Comets are often divided into two groups, short- and long-period comets. The short-period comets have orbital periods less than 300 years and have orbits that are quite aligned with the planetary plane, whereas the long-period comets have randomly oriented orbits and may have orbital periods up to a million years.IN 1950 Jan

Oort made a prediction that the long-period comets originate from a spherically shaped region surrounding the solar system, now named the Oort cloud. The Oort cloud has never been directly observed but the orbits of the long-period comets strongly suggests its presence. The comets in the Oort cloud were probably not formed there but were rather catapulted outwards by the Jovian giants. In 1951 Kuiper proposed that another closer and disk shaped compound of comets must exist where the short-period comets can originate from. This is called the Kuiper belt and lies between 30-100 AU. Kuiper belt objects has been observed.

3 Observed Molecules

The composition of the nucleus of a comet is still rather unknown. Measurements have so far been restricted to the coma. When the comet approaches the sun the ices begin to evaporate creating the large coma which can be about 30 times the size of the earth. This is very thick compared to the small (about 5 km) nucleus. Fluorescence and reflected light from the sun diminishes our chances of a clear view of the nucleus. When the comet is far away enough from the sun so that the coma has disappeared, it is also to far away to be resolved and analyzed. So until we have actually landed on a comet and taken direct samples, all the information about the comets composition comes from the coma. But as the ices evaporate and expands it will undergo chemical processes, mostly due to UV radiation from the sun. Therefore one must guess what kind of original molecules there where that could result in the coma composition. The solutions are however imprecise and not unique. Furthermore the components of the coma originates from two different sources. Mostly of course from evaporated ices from surface of the nucleus but also from disintegrated dust particles that are released, as the ices evaporate. The surface of the nucleus is probably also not representative for the whole nucleus. The elemental abundances can be established with fairly good accuracy, and give additional clues to the molecular abundances.

Element	Abundance
H	0.5464
C	0.1137
N	0.0132
O	0.2834
Na	0.0010
Mg	0.0099
Al	0.0007
Si	0.0183
S	0.0071
Ca	0.0006
Cr	0.0001
Fe	0.0052
Ni	0.0004
Total	1.0000

The dust particles can be both of non-organic and organic nature based on so called CHON (carbon, hydrogen, oxygen, nitrogen). It is difficult to determine the mass-ratio

between the released gas and dust because of the small number of large particles. Small particles may be difficult to detect. The ratio lies typically between $0 < X < 2$.

All molecules identified in the coma are also found in the interstellar medium except S_2. There are ions and radicals however that are not observed in the ISM. This depends on the facts that the solar radiation gives rise to photo processes. The similarities between molecular presence in comets and interstellar clouds should not be over interpreted, the relative abundances may differ largely, due to how the comets where formed. The organic molecules are mainly in the dust particles, the exact molecules are still unknown but several complex forms such as polymers must be present. One of the six essential elements for life (H, C, N, O, S, P), phosphor has not been detected in comets, however if its relative abundance is as low as in the sun it may just be difficult to detect. More than 70 different amino acids has been found in meteorites which strongly suggests that they might be present in comets too. What can be said about observations are that the Whipple dirty snowball model has been confirmed to a high degree. Water is the dominant ice, although the exact ratio is difficult to determine, more volatile types may have evaporated earlier on the comets orbits towards the inner solar system. Very volatile molecules and elements are not observed in comets, if they could condense and thus could be incorporated into the comets during formation or if they have evaporated later on is not clear. Carbon monoxide, carbon dioxide and methanol have high abundances around 5% each. This is also the case for formaldehyde, but is expected to be a secondary product from other molecules. This is a conclusion drawned by the fact that formaldehyde seems to be more abundant further out in the coma. Polyoxymethylene, the solid polymer of formaldehyde could be the answer, the process for this is studied by Herve Cottin, Yves Benilan, Marie-Claire Gazeau and Francois Raulin. They conclude that so far this is the best explanation, there is no known cometary gaseous compound that can explain the observations. The existence of polymers is important in the hypothesis that complicated organic molecules is present in comets. Around 1% is ammonia and methane respectively, both important in the process of amino acids. Lots of different CHO molecules has also been observed such as formic acid, acetaldehyde and methyl formate. The findings of these molecules increases the hope that even more complicated molecules may be present. Acetylene and ethane also suggest that lots of different hydrocarbons might be present. Hydrogen cyanide(HCN) has also been detected in comets. It is an important molecule for prebiotic chemistry. It also seems that HCN gets more abundant in the coma as the comets approach the sun. This suggests that it may originate from organic grains, possibly in the form of polymers, which can be a source for amino acids. Hydrogen sulphide is also an important constituent with an abundance around 1%. The total weight % of organic molecules in comets is estimated to be between 20-50%?.

Molecule	Name	Abundance(relative to water)
H_2O	Water	100
CO	Carbon monoxide	2-20
CO_2	Carbon dioxide	2-6
CH_4	Methane	0.6
C_2H_6	Ethane	0.3
C_2H_2	Acetylene	0.1
H_2CO	Formaldehyde	0.05-4
CH_3OH	Methanol	1-7
$HCOOH$	Formic acid	0.1
$HNCO$	Isocyanic acid	0.07
NH_2CHO	Formamide	0.01
CH_3CHO	Acetaldehyde	
$HCOOCH_3$	Methyl formate	0.1
NH_3	Ammonia	0.5
HCN	Hydrogen cyanide	0.1-0.2
HNC	Hydrogen isocyanide	0.01
CH_3CN	Methyl cyanide	0.02
HC_3N	Cyanoacetylene	0.02
N_2	Dinitrogen	0.02-0.2
H_2S	Hydrogen sulphide	0.3-1.5
H_2CS	Thioformaldehyde	0.02
CS_2	Carbon disulphide	0.1
OCS	Carbonyl sulphide	0.4
SO_2	Sulphur dioxide	0.2
S_2	Disulphur	0.05

3.1 Carbonaceous Chondrites

Clues to the composition of comets may be found in a special kind of carbon rich meteorites called carbonaceous chondrites. Although most meteorites originate from the asteroid belt, the origin of the carbonaceous chondrites is under debate. Since their composition is different from other meteorites a different source might be an option. One proposal is that they are comet drop outs from comets close to their perihelion. Whatever their origin might be there is no reason to rule out the possibility that their composition could not be similar to that that in comets. If amino acids are incorporated in them when they are created or formed in them, this would be a probable case also for comets. And amino acids have been found, in fact lots of them. More than 70 different amino acids has been observed in these chondrites, and 8 of them overlap with the 20 used by life on earth. This is a strong evidence that amino acids important to life could be present in comets as well.

3.2 Comparison with the Interstellar Medium

How closely linked are the comets to the interstellar medium(ISM)? There are competing theories in this area. It is argued that comets could not have been formed directly from the interstellar medium, accreting mechanisms in the ISM could not explain such big objects

to form and during the formation of the planetary disk it is not likely that the comets could have survived unprocessed. Advocaters of this model suggest the above description of formation of comets(sect 2). But there are observational evidence that could point in another direction. First it is interesting to notice that most molecules found in the comets are also found in the interstellar medium, except CO_2, S_2 and the radicals NH_2 and NH. CO_2 lack dipole moment and have no rotational spectra. But the presence of CO_2H^+ suggest that CO_2 also is present. NH_2 and NH have probably not been observed because their rotational spectra lies in the sub millimeter region where observational method so far have been poor. Recently however there has been indications for NH_2 so NH is probably present to. The only real mystery here is S_2, which like CO_2 lacks rotational spectra but should have been observed through its fine-structure transition. There are also many ions present in the coma that are not detected in the ISM, which is not surprising since the coma is subjected to high radiation from the sun. The ratio between deuterium and hydrogen in water and HCN gives also important evidence. For three different comets this ratio has been established to be 0.0003, this is about 10 times the expected value for a protosolar nebula and about twice the value in the oceans of the Earth. Isotopic ratio increasing is mainly an effect of chemical processes between ions and neutrals at low temperatures. This probably took place in the interstellar medium and suggests that if these molecules survived the formation of the solar system than so could others. There is a close similarity between interstellar grains an comets. Not only in what kinds of compounds that are present, but also in their relative abundances. This is true for many of the organic species. They also share molecules of silicates and a high abundance of metals. The similarities between interstellar grains has led people among them J.M. Greenberg and his team to conclude that cometary nuclei is accumulated interstellar grains. This is not totally in accord with the model of formation of comets presented earlier. However some grains could have survived the formation of the presolar nebula, to be incorporated with comets later. It is also possible that grains like those in the ISM could be formed later on in the presolar nebula. There seems however to be a close link between the interstellar grains and the comets, it is therefore useful to study the ISM in order to get clues to what molecules that can be expected in comets. There are over 130 different molecular species observed in the ISM many of organic nature and prebiotic importance. Lately there has been evidence of glycine, the simplest of the amino acids, in hot molecular clouds. Glycine is a possible parent molecule in comets. Other molecules with biochemical importance, apart from those also found in comets, include ketene, vinyl alcohol, acetetic acid, glycol aldehyde, ethanol, acetone, and benzene. For some of these, like benzene, it is indicated by mass spectroscopy that they also exists in comets.

4 Evolution of Comets

New research suggests, to complicate the interpretations of observations further, that there are evolutionary processes that occurs in the comets during the long periods they spend in the Kuiper belt and Oort cloud. These can be divided into thermal, collisional, radiation and interstellar medium processes. The result is that future samples from cometary surfaces may have been modified substantially from the time when the solar system was born. Thermal processes occur when stars pass by closely or a nearby supernova explosion take place. A passing star can heat the Oort cloud well above the usual 5-6K, for example in the last 4.5Gyr it is almost certain that some O type star heated the Oort cloud up to 16 K which would remove condensed neon and molecular oxygen. Passing stars have little

effect on the Kuiper belt however, that have a temperature between 30-60K. About 30 supernova event during the last 4Gyr are estimated to have heated the surfaces of Oort cloud objects to above 30K maybe some even as high as 60K. At 30K CO, N_2, and CH_4 would leave the surfaces, at 60K also formaldehyde. Since supernova explosions are rather brief this effect will not penetrate deep into the comets, only 1-2m, the passing O stars may effect on much deeper levels, up to maybe 50 m. Collisions in the Oort cloud are rare due to the large distances between objects and their low rotation speeds. In the Kuiper belt on the other hand collisional process might be important. The space between objects and their rotational speed makes collisions much more frequent and severe, causing comets to lose surface material over time. In fact collisions are so frequent that it is now believed by some that all present day comets in Kuiper belt might actually be fragments lost by larger objects in collisions.

When the solar system passes through giant molecular clouds the interaction between the molecules and comets causes the comets to shed surface material. In the Oort cloud this process could have taken up to 20 meters of surface material.

Ultraviolet solar and interstellar photons and high energetic charged particles provide enough energy to break up molecules and start chemical processes on the surfaces. Experiments have been made that suggest polymerisation of molecules in comets under UV radiation. This leads among other things to long chained hydrocarbons that darken the surface.

Though all these evolutionary processes, comets are still the most pristine objects known, and do not lose their role as information carriers from the early solar system. But evolution must be taken into account when interpreting cometary observations.

5 Survival Rate of Organic Molecules in Comet Impacts

In 1989 it was reported that extraterrestrial amino-acids had been found above and below Cretaceous-Tertiary boundary. This was taken as a proof that the amino-acids had formed in the post impact shock (explained below), since there were not present in the boundary layer itself. Later extraterrestrial amino acids have been found in the boundary suggesting that organic material can survive large impacts. One of the main arguments against organic molecules arriving from cometary impacts is that the temperature in the collision would be too great for the molecules to survive. The minimum relative velocity for an impact is 11.2 kms^{-1}, which would give rise to huge temperatures, of the order of 10^4K, that could break up organic molecules. But heterogeneity in the explosion could leave some regions with lower temperature. There have been reports of extraterrestrial He trapped in fullerenes that seems impossible if the fullerenes was exposed to temperatures above 1000^oC. This also indicates that organic molecules can survive impacts.

5.1 Experimental Results

To test what happens to amino acids when shocked, Jennifer G.Blank and her team has conducted a series of experiments. To simulate a cometary impact a gun fired a can sized bullet at a small metal device containing a drop of water mixed with amino acids. A large fraction of the amino acids survived the shock, and also very interestingly did some of them form chains of two, three and four amino acids. This is on the way to form proteins which are made of longer such chains. If the container was freezed even more amino acids

survived. The experiment simulated low angle impacts, less than 25 degrees in order to decrease the severity of the impact, and with the hypothesis that jets are created lifting some material away from the highest temperature. At such low angles some of the water could maybe remain intact and result would be a pond of organic material and water, an ideal birthplace for life. Analysis of the shocked amino doped fluids showed a survival rate of 40-75%.

Earlier experiments used high-power pulsed laser to create high temperature shock waves, in a CH_4 rich gas, to show how a shock can synthesize organic material was conducted by Christopher P. McKay and William J.Borucki. The initial plasma temperature were over 10000K but the gas rapidly cooled as the shock propagated outwards. Two regions of the gas could be identified. An inner region where large temperatures and pressure were present, here all molecules had time to dissociate. But there was also formed an outer cooler where the more stable molecular species could survive. Organic material was also produced by exposure by UV lightning and/or chemical reactions with cooler gas. UV light may cause CH_4 together with other elements to form different prebiotic compounds, in regions where the shock is to weak to destroy them. McKay and Borucki argues that a comet impact is a situation where hot gas is suddenly introduced in a cooler atmosphere. The result is a shock wave propagating outwards engulfing surrounding gas, which is then heated by mixing and radiation. The early earth might be rich in CH_4, (from comets maybe), which could be synthesized into more complex molecules by shocks from further comets or maybe by lightning. Thus even if lots of the molecules are initially destroyed by the impact, further shocks can build new organic material.

5.2 Impact Simulations

There have also been sophisticated computer simulations of comet and asteroid impacts. E. Pierazzo and C.F. Chyba presents extensive works in this area, where different impacts have been studied. Among others comet versus meteorite impacts were analyzed. In the comet impacts the material from the comet enters the expansion plume and is cast away from the impact area early and can cool down. In the asteroid case the material remains in the crater for a longer time and experiences higher temperatures and pressures reducing the chances for organic molecules to survive. Other parameters that were investigated were the velocity and size of the impacting comet. Not surprisingly higher speeds resulted in bigger shocks and higher temperatures, for larger objects the material is shocked for a longer time. All these effects reduces the survivability. The calculations give a 0.1-10% rate of survival for different amino acids for a 1km body coming in vertically at a speed of 15 kms^{-1}. For a 20km sized object at 20 kms^{-1} only 0.02 percent of the most stable amino acids survived. The impact angle is also an important factor , vertical impacts have very low survivability of amino acids, but for some amino acids the rate of survival is increased by 3-5 times when the angel is shifted to 45^o, the most probable angle of impact. Inhomogeneities can also create jets during the impact redistributing temperature and material and leaving some parts cooler. Although the rate of survival of amino acids in most cometary impacts is low they would for a long time deliver amino acids to the earth oceans, and the occasional low angle impact would create large concentrations on various areas. For an impact angel of 5^o tens of percent of certain amino acids could survive according E.Pierazzo and C.F. Chybas simulations. These concentrated areas of organic material including amino acids could be ideal places for life to start. This is also one of main advantages with the theory that life was ignited from a cometary impact. Other models such as Urey-Miller and infall from interplanetary dust particle can provide

a steady increase of amino acids but no large concentrations.

5.3 Infall Rate

Chyba and Sagan estimated the infall of organic molecules from comets to $10^{11} kg\ yr^{-1}$ during the heavy bombardment era. These results are based on the amount of organic material found in comets and the rate of impacts calculated from the number of craters on the Moon. Other mechanisms for bringing organic matter to earth include infalling dust particles and terrestrial sources such as UV photolysis. These are estimated by the same authors to be of the order of $10^{8-12} kg\ yr^{-1}$ for terrestrial sources depending on the atmosphere and $10^{8} kg\ yr^{-1}$ for in falling dust. These results are very uncertain and the authors themselves calls for future revision when further knowledge is obtained about comet composition, survival rate and the early atmosphere.. It does however suggest that comets can have provided a substantial amount of the organic material on the early earth.

6 Future Observations

As we have seen, observations have so far not been enough to completely determine the compositions of a comet. The elemental abundances together with some identified molecules and the dust-ice ratio is all we have so far. But in order to know the exact molecules and by so further investigate comets role in the origin of life, an actual landing on a comets nucleus is necessary. On the second of march 2004 such a mission was launched. The mission is called Rosetta after the Rosetta stone that helped provide a key to understand the Egyptian hieroglyphs. The prime goal of the Rosetta mission is to provide information about cometary material in order to find a relationship between cometary and interstellar material. This will hopefully give clues about the origin of the solar system. But it will also give information about what kind of organic molecules that are present in the comet nucleus. On the way to the comet Rosetta will also fly by and examine two asteroids. The target comet is called 67P Churyumov-Gerasimenko. This is a comet with quite an unusual history, prior to 1840 it's perihelion distance was 4.0 A.U. and unobservable from Earth. It was then disturbed by Jupiter causing it to move inwards, and then a new encounter with Jupiter in 1959 caused its perihelion distance to shrink down to only 1.29 A.U. The Rosetta probe will follow the comet all the way from far out close to its aphelion to its perihelion, to explore how the cometary activity evolves. It will try to identify volatile material along the way. On board the orbiter there are cameras, spectrometers for ultraviolet, visible and infrared, and also a radio telescope. There are also mass spectrometers and dust analyzers for a direct measurement of the coma, which it will be in. Earlier mass spectrometers used in comet flybys has not had the resolution needed, for example molecules with the same weight but different atomic numbers has not been possible to separate. The lander also carries a mass spectrometer but is also equipped with X-ray and gamma-ray spectrometry. Probes for measuring temperatures at different depths and times are on board so that the thermal properties of the nucleus can be determined. The Rosetta mission will certainly provide us with much needed information. Another ambitious mission was launched in 1999 and flew by comet 81P/Wild 2 in January this year (2004) and collected dust from the coma which it will return for direct analysis on earth in 2006.

7 Life in Comets

It has been suggested by Fred Hoyle among others that not only organic molecules arrived from comets and asteroids but also ready life forms. This tries to explain how life could emerge so quickly in earth's history. These life forms would maybe have been ejected from planets and later incorporated in comets. Others suggest that life actually formed in comets. We have seen evidence for large amounts of water and organic molecules so why not? A problem these theories face is that life is in need of liquid water. An idea is that as the comet approach the Sun, a layer of melted water forms in the surface areas, this is the only region where the sun is expected to be able to do so. But in the surface areas the microbes would also be targeted by UV light that would have seriously damaging effects. A solutions is proposed to be internal heating through radioactive decay, and especially the decay of ^{26}AL. This is an isotope that has been observed to have a steady production rate in the galaxy, through supernovae and maybe novae explosions and should have been abundant during the formation of the solar system. Models calculating this suggests that comets larger than 10-15km should have been able to sustain liquid cores for about 10^6 years, the presence of other decaying material could maybe prolong this. When freezing again a hollow core would be created. Wallis conductor of these models, proposed that this also is the reason why large comets break up so easily. The fluid core would be quite an ideal place for life to start. However more recent studies that includes trapped gas in the model give a more pessimistic response to the idea of liquid water. Our knowledge of the composition of comets is still to poor for these models to provide certain answers.

Micro-organisms have been found on earth that has survived 3 million years of hibernation in a temperature of -10^oC. There is even some that are suspected to have been hibernating for up to 100 Myr. So it is possible for freezed bacterias to be carried around on comets or in dust grains. In the comet case these would be shielded from cosmic radiation to some extent but for those in grains the radiation could pose a problem. Relating hibernated bacterias on earth that has survived for between 0.5 and 100 Myrs subjected to radoactivity $1 - 10mGyyr^{-1}$ to supposed bacterias in space gives a hopefull answer. The radiation in space would be around $100mGyyr^{-1}$, which would mean that micro organisms would accumulate deadly doses on a timescale around 1-10Myr.
The impacts is another matter, for panspermia to work the transporting body must not create a too large impact. Freeze-dried bacterias, which is a likely state for space bacterias, can survive flash heatings up to 350^oC for a period of 30 seconds. Temperatures well above this occurs in impacts but under lucky circumstances maybe bacteria could survive. It is thus proposed by some that life started on another planet circling another star, that after an impact was thrown into space. These would later be incorporated into the interstellar cloud that formed our solar system, frosen into comets and later arriving on earth. Models for the formation of our solar system predicts that the temperatures in the outer parts during formation would not reach above levels that would destroy these organisms. That boulders are ejected into space is clear, it is estimated that about 15 marsian meteorites each year land on earth. It is however unclear how the ejected material could escape the parent star and float into interstellar space. W.M Napier suggest that as metre-sized boulders are ejected into space they are eroded into small particles by collision of zodiacal cloud particles (small particles in our solar system, such as dust from comets) . Small particles can then be ejected into space from the solar system by radiation pressure. Napier estimates that 10^{20} such particles per million years is ejected out of the solarsystem. When the solar system passes through dense molecular cloud these particles would be incorporated with them and may be a part in the formation of a new

solar system. So maybe not only life arrived from space but the earth also seeded other planets. Hoyle and Hoover claim to have achieved spectroscopic evidence for diatoms, a small algae in space. The absorbance of the interstellar medium at 2200, is due to the fact that interstellar grains main constiuent is frozen microbes they claim. These result are highly controversial. It is not generally believed that interstellar grains have a large fraction of microbes, but rather a mixture of crystaline and normal carbon. It has also been pointed out that small particles ejected by radiation pressure incorporated into giant molecular clouds would also be subjected to erosion by the cloud and probably would be totally eroded in this process. Observations contradict this however, to explain the observed extinction, the interstellar medium must contain large enough particles that could contain bacteria. The organisms need not directly be lifted by radiation pressure. Escaped bacteria from planetary ejecta could also be incorporated into long period comets, travelling far awy from the solar system. This would be a more shielded environment. When transversing molecular clouds material would be shedded and incorporated. In dense H II regions, in which grains has been observed to radiate at 300K, the expected life time of a comet is of the order of a few thousand years. This means that the entire comets would be dissolved in the outskirts of the cloud, and during star formation be included in the planetary system. Through comets, direct delivery to protoplanetary discs and ready planetary systems is also possible. Although chances of direct delivery is small a Jupiter like planet could disturb the comets orbit and/or make it break up, which makes delivery to to the Jupiter like planets satellites a little more probable. The estimated interval between protoplanetery discs flybys close enough is 1 in every 1.5Gyr and 1 every 25Myr for GMCs. The delivery to the protoplanetary discs is also much smaller, a 5-kg can be estimated, whereas a around 2 tonnes is expected to be delivered to a GMC. The GMCs therefore receive a lot more, but the survival of the microbes during star and planet formation is uncertain. A few kg is not insufficient to spawn life, perhaps incorporated into comets during planetary formation, where the interior is heated by radioactive decay and the bacterias could multiply before delivery to planets.

8 Summary and Conclusions

Complex organic molecules are found in comets and in the interstellar medium. In carbonaceous chondrites more than 70 different amino acids have been discovered. This along with experiments and simulations showing that organic molecules can survive an impact providing the earth with large concentrations of building blocks for life, strongly indicates that this model for providing earth with means to create life should not be ruled out. Not only is the calculated infall comparable to or greater than terrestrial sources but comets also provide a very good way for creating ponds of water with a high concentration of organic material when the occasional low angle impact on a rocky surface occurs. Further observations are needed to gain a more thorough understanding of the composition of comets and what kind of complex organic molecules are present. The lack of knowledge concerning the atmosphere of the early earth also poses a problem in understanding the comets role versus organic sources on earth. Until we gain further knowledge, much of this science is speculations although with much reason behind them. The panspermia theories, far-fetched as they seem, should not be ruled out either; mechanisms for transportations of bacteria from planet to planet and solar system to solar system have been explained to some degree although much uncertainty remains.

Acknowledgements

Thanks to Anders Winnberg for valuable correspondence.

References

Blank J.G., Knize M.G., Nakafuji G., 2002, Lunar and Planetary Science 33, 2075.

Blank J.G., Miller G.H., Ahrens M.J., Winans R.E., 2001, Origins of Life and Evolution of the Biosphere 31, 15.

Botta O, Glavin D.P., Kminek G., Bada J.L., 2002, Origins of Life and Evolution of the Biosphere 32, 143.

Chyba C.F., Sagan C., 1997, Comets and the Origin of Life (New York, Berlin: Springer), 146.

Cottin H., Benilan Y., Gazeau M-C., Raulin F., 2003, Icarus 167, 397.

Crovisier J., Encrenaz T., 2000, Comet Science (Cambridge University Press) 117.

Ehrenfreund P., Irvine W., Becker I., Blank J.G., Brucato J.R.,Colangeli L., Derenne S., Despois D., Dutrey A., Fraaije H., Lazcano A., Owen T., Robert F., 2002, RepProgPhys 65, 1427.

Huebner W.F., 2002 Earth Moon and Planets 89, 179

Huebner W.F., Boice D.C., 1997, Comets and the Origin of Life, (New York, Berlin: Springer), 111.

Kissel J., Krueger F.R., Roessler K., 1997, Comets and the Origin of Life, (New York, Berlin: Springer), 69.

Kuan Y-K., Charnley S.B., Huang H-C., Kisiel Z., Ehrenfreund P., Tseng W-L., Yan C-H., 2004, Advances in Space Research 33, 31.

McKay C.P., 1997, Comets and the Origin of Life, (New York, Berlin: Springer), 273.

Mckay C.P., Borucki W.J., 1997, Science 276, 390.

Napier W.M., 2004, MonNotRAstronSoc 348, 46.

Pierazzo E., Chyba C.F., 1999, Meteoritics and Planetary Science 34, 909.

Podolak M., Prialnik D., 1997, Comets and the Origin of Life, (New York, Berlin: Springer), 259.

Stern S.A., 2003, Nature 424, 639.

Svetsov V.V., 2002, Solar System Research 36, 50.

Wallis M.K., Wickramasinghe N.C., 2004, MonNotRAstronSoc 348, 52.

Student Workshop

Hot Topics in Astrophysics 2003/2004

The astrophysics students welcome you to a workshop in modern astrophysical research

Thursday 6 May

Room N6115, Origo building, 6th floor, Physics

9:30-12:15

- Science with the New Sub-Millimetre APEX telescope
- Wavelets in Radio-Astronomy
- Dark Matter in the Universe and Alternatives
- Probing the Acceleration of the Universe
- Dark Energy and Quintessence in the Universe

13:15-15:30

- Starburst in Merging Galaxies
- Extrasolar Planets
- Life in the Universe
- Organic Molecules in comets

Contact: Alessandro Romeo (romeo@fy.chalmers.se)

STUDENTS' WORKSHOP
"Hot Topics in Astrophysics 2003/2004"

Thursday 6 May 2004

Room N6115 (at Chalmers, Physics, Origo building, 6th floor)

ORGANIZER: Alessandro Romeo (romeo@fy.chalmers.se)

PROGRAMME

- **09:30–09:35**
  Welcome to the audience
  — Alessandro Romeo

Session 1 (Chairperson: Mats Johansson)

- **09:35–10:05**
  Science with the New Sub-Millimetre Telescope APEX
  — Christophe Risacher
- **10:05–10:35**
  Wavelets in Radio-Astronomy
  — Rajat Mani Thomas

COFFEE BREAK

Session 2 (Chairperson: Markus Janson)

- **10:45–11:15**
  Dark Matter in the Universe and Alternatives
  — Farhad Aslani
- **11:15–11:45**
  Probing the Acceleration of the Universe
  — Martin Nord
- **11:45–12:15**
  Dark Energy and Quintessence in the Universe
  — Daniel Johansson

Session 3 (Chairperson: Farhad Aslani)

- **13:15–13:45**
  Starbursts in Merging Galaxies
  — Raquel Rodriguez Monje
- **13:45–14:15**
  Extrasolar Planets
  — Markus Janson

COFFEE BREAK

Session 4 (Chairperson: Rajat Mani Thomas)

- **14:25–14:55**
  Life in the Universe
  — Mats Johansson
- **14:55–15:25**
  Organic Molecules in Comets
  — Niklas Vahlne
- **15:25–15:30**
  Thanks to the audience
  — Alessandro Romeo