Universe: Simulations of Structure and Galaxy Formation

The current cosmogonic paradigm posits that structures in the universe (such as galaxies and clusters) originated from tiny density fluctuations generated around the time of the Big Bang and subsequently amplified by gravity. The precise origin of the primordial density fluctuations is uncertain. One possibility is that they were generated when the universe underwent a brief period of exponential expansion known as inflation, a tiny fraction of a second after the Big Bang. In the most successful model, the mass density of the universe is dominated by an exotic form of matter called cold dark matter which consists of elementary particles that do not make any contribution to the luminosity density of the universe. The predictions of this model have been set out extensively, using both analytic calculations and computer simulations, the latter having become increasingly important over the past 20 years. The build-up of structure in a cold dark matter universe is hierarchical: small-mass objects are the first to condense out from the expanding universe at early times (high REDSHIFT), whilst more massive objects (such as rich clusters) form later by repeated mergers of smaller objects. A remarkable confirmation of this framework was provided by the discovery, in 1992, of anisotropies in the temperature of the COSMIC MICROWAVE BACKGROUND radiation (CMB) by the DMR instrument aboard the COBE satellite. The amplitude of these anisotropies was within a factor of two of the extant theoretical predictions. Spurred on by the deep images of the sky produced by the Hubble Space Telescope, which show galaxies as they were when the universe was only about 10% of its present age, a great deal of effort is now being directed towards modeling the formation and evolution of galaxies within the setting of cosmological structure formation in the dark matter. The dissipative gas dynamical processes involved in GALAXY FORMATION make this a challenging task that can only be tackled using either idealized models or large amounts of supercomputer time.

The physics of galaxy formation

Studies of the dynamics of galaxies confirm a fundamental inference made 25 years ago from studies of the kinematics of stars within galaxies: on scales larger than galactic nuclei, the dominant physical interaction is gravity. Furthermore, it is now well established that the predominant source of gravity is dark matter, that is matter that does not emit detectable electromagnetic radiation. Dark matter is now routinely 'imaged' through the phenomenon of GRAVITATIONAL LENSING, the relativistic deflection of the light from distant galaxies as it passes near an intervening cluster of galaxies. Independently of the exact identity of the dark matter, the dominance of gravity leads directly to a general scheme for structure formation, first outlined by Landau and Lifshitz in the 1950s, rigorously developed by Peebles during the 1970s, and calculated in detail using computer simulations in the 1980s and 1990s. The key concept is the gravitational instability experienced by small matter overdensities in the expanding universe. Matter fluctuations present in the early universe grow in amplitude approximately as a power-law in time and eventually collapse to form self-gravitating objects.

The process of gravitational instability sets the scene for galaxy formation, the main physical ingredients of which are set out in figure 1. Although the precise details depend on the identity of the dark matter, under quite general conditions, galaxy formation is expected to proceed via a two-stage process originally outlined by White and Rees in 1978. First, gravitational instability acting on collisionless dark matter, results in the formation of self-gravitating dark matter haloes. Gas, initially well mixed with the dark matter, participates in this collapse, but it is heated by shocks to the thermal (or virial) temperature of the dark matter haloes. Second, the hot gas cools radiatively, on a time scale set by atomic physics, due to bremsstrahlung, recombination and collisionally excited line emission. The rate of cooling depends upon the density of electrons and atomic nuclei, and so it was most efficient at high redshift, when the universe was denser. (In practice, the heating and subsequent cooling of the gas may occur rapidly and chaotically, particularly in small galaxies.) Just prior to gravitational collapse, angular momentum is imparted on the generally aspherical perturbations by gravitational torques exerted by neighboring clumps, as proposed by Hoyle in 1948. Thus, the initial collapse generically results in the formation of a gaseous disk. Once the disk has become centrifugally supported, the material in it begins to fragment into stars by processes that are still poorly understood. In this simplified picture, the spheroidal components of galaxies form by mergers of disk galaxies which jumble up the stellar orbits, disrupting their organized configuration.

The scheme outlined above provides a natural explanation for why there are galaxies of two basic types: disks and spheroids. It also elegantly explains why there is a limit to the luminosity that galaxies attain at the present day. The most luminous galaxies form in the most massive dark matter haloes. These have only recently collapsed because, as is the case in most current models, the amplitude of mass fluctuations is smallest on large scales. At the high temperatures and low gas densities prevailing in recently formed structures, cooling is very inefficient and the gas has not had time to cool and fragment into stars over the lifetime of the universe. One problem of hierarchical structure formation models is that the number of low-mass haloes that form exceeds the number of faint galaxies seen in the local universe. However, a likely solution is that the feedback of energy into the cooled gas from early generations of stars will have acted as a self-regulating mechanism. This feedback prevents substantial star formation activity in shallow gravitational potential wells, thus causing most of the small-mass haloes to harbor extremely faint galaxies.

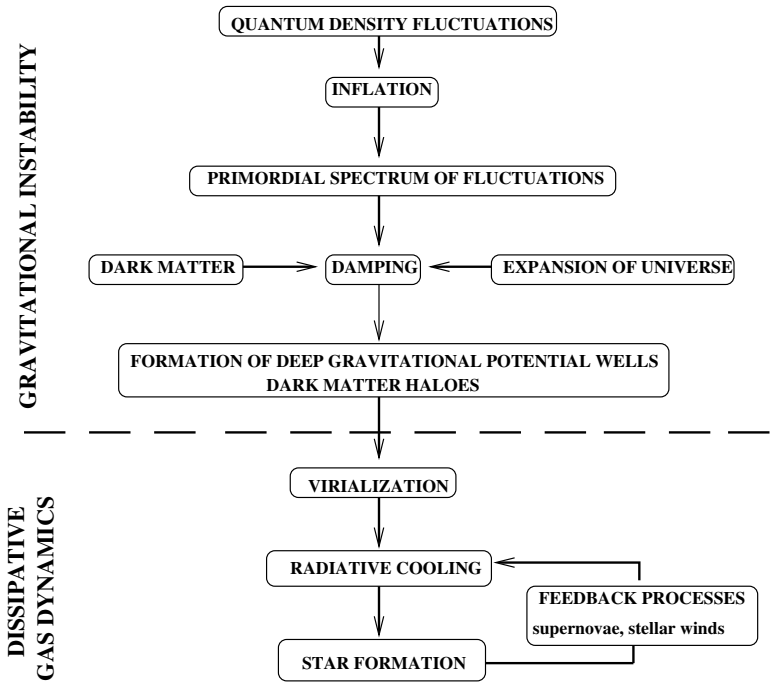

Figure 1. The ingredients of the standard model for the formation of galaxies and cosmological structure.

Cosmological structure formation

The processes of gravitational instability and collapse, gas cooling and star formation operate under quite general conditions. A quantitative theory of galaxy formation, however, requires that two key cosmological questions be addressed: (i) what is the origin of primordial mass fluctuations and (ii) what is the identity of the dark matter? Significant progress towards answering these questions was made in the early 1980s, as the result of a fruitful interaction between particle physics and cosmology.

The first influential idea of the 'New Cosmology' was proposed around 1980 by Alan Guth and extended by Andrei Linde. Searching for a solution to the 'magnetic monopole problem', Guth proposed that the universe had undergone a period of exponential expansion, or INFLATION, very soon after the Big Bang, triggered perhaps by the transition of a quantum field from a false to the true vacuum. Quantum fluctuations generated during this epoch would be swept across the event horizon and thus become established as classical ripples in the energy density of the universe. When they cross the horizon back again, their amplitude is independent of scale. Thus, the inflationary model predicts a scale-invariant power spectrum of primordial fluctuations, $P(k) \propto k$, with a Gaussian distribution of amplitudes. (More elaborate versions of the same idea can produce models whose power spectra have an exponent that differs slightly from unity.) The subsequent evolution of the fluctuations depends on the values of the cosmological density parameter, Ω_0, the Hubble constant, H_0, the COSMOLOGICAL CONSTANT, Λ_0, and the identity of the dark matter.

The second key idea from the 1980s concerns the identity of the dark matter. The abundance of the light elements (H, D, He, Li) produced during Big Bang nucleosynthesis agrees with astronomical data only if the present-day density of baryons is low enough for deuterium to form in at least the abundance measured in primitive gas clouds at high redshift. The baryon density required for this is about one order of magnitude smaller than the total mass density of the Universe inferred through a variety of tests. Thus, a fundamental conclusion is that the dark matter must consist of non-baryonic elementary particles.

Particle candidates for the dark matter are conveniently classified into 'hot' and 'cold' varieties, a nomenclature introduced by J R Bond around 1980. The prototype of a hot particle is a stable neutrino with a mass of the order of a few eV. (A single species of neutrino would give $\Omega = 1$ if its mass were approximately 30 eV and $H_0 = 50$ km s^{-1} Mpc^{-1}.) Examples of cold particles include the least massive stable supersymmetric particle, the neutralino, and the much lighter axion. Cold particles are often referred to as weakly interacting massive particles or WIMPs. There is a fundamental difference in the way in which galaxies are predicted to form in hot and cold dark matter cosmogonies, arising from the different damping mechanisms that affect the development of primordial fluctuations in the two cases. If the universe were dominated by massive neutrinos, then fluctuations below

some critical mass would be wiped out because the neutrinos move at relativistic speeds in the early universe and rapidly free-stream out of overdense regions. For a single neutrino of mass 30 eV, this critical scale corresponds to a fluctuation wavelength of

$$\lambda_c = \frac{2\pi}{k_c} = \frac{13}{\Omega h^2}\ \mathrm{Mpc} \qquad (1)$$

where we have parametrized Hubble's constant as $H_0 = 100h$ km s^{-1} Mpc^{-1}. The power spectrum decays exponentially for wavenumbers $k > k_c$.

In the case of cold dark matter, the damping of density fluctuations is much less severe. The free-streaming scale of cold dark matter is many orders of magnitude smaller than that of hot dark matter, and so this effect is not important. The most relevant damping process is the Mezaros effect, whereby oscillations in the radiation energy density stifle the growth of matter fluctuations. This situation persists until the energy density of matter dominates over that of the radiation, which occurs after a redshift of $1 + z_{eq} = 23\,900\Omega h^2$. The size of the horizon at this epoch is imprinted upon the cold dark matter power spectrum, marking the turnover of the primordial form, $P(k) \propto k$, to the damped fluctuation spectrum which asymptotes to $P(k) \propto k^{-3}$ at small scales.

In neutrino dominated models, the first structures that form are flat, pancake-like objects of mass comparable to that enclosed within the critical free-streaming scale, corresponding approximately to $10^{16} M_\odot$. These are objects of supercluster scale which must somehow subsequently fragment, in a 'top-down' fashion, in order for galaxies to form. Early computer N-body simulations of this process carried out by Frenk, White and Davis in 1983 showed that for a neutrino dominated universe not to exceed the level of clustering measured in the galaxy distribution today, galaxies would need to form at redshifts $z \lesssim 1$. Yet, it was already known at that time that quasars can have much higher redshifts than this and, today, we know that there is a large population of galaxies already established at $z = 3$–5. Thus, although rather appealing at first sight, neutrino dominated universes with Gaussian primordial fluctuations were soon abandoned.

The alternative, a cold dark matter universe, proved to be much more successful, as discussed, for example, in a series of papers in the 1980s based on N-body simulations, by Davis, Efstathiou, Frenk and White. The defining property of the fluctuation spectrum in a cold dark matter universe is that small-scale perturbations are preserved. Thus, subgalactic mass haloes are the first to collapse and separate out from the expansion of the universe. These haloes then grow, either gradually by accreting smaller clumps, or in big jumps by merging with other haloes of comparable size. The timetable for the formation of structure in a universe dominated by cold particles is hierarchical or 'bottom-up'—small objects form first, larger objects form later. The cold dark matter fluctuation power spectrum thus specifies completely the evolution of the merging hierarchy of dark matter haloes into which the baryons must fall in order to make the galaxies.

Currently, the most successful version of the cold dark matter model has around 30% of the critical density in cold matter and 70% in the form of a vacuum energy density or cosmological constant term, so that the universe has a spatially flat geometry. This model is in good agreement with a range of observational data: the amplitude of the angular power spectrum of temperature fluctuations in the cosmic microwave background (including the location of the first 'Doppler' peak), the abundance of clusters of galaxies ranked by their x-ray luminosity, the clustering of galaxies on large scales, and the expansion rate of the universe as deduced from the brightness of distant supernovae.

The recent detection of oscillations in neutrino flavour by the Super-Kamiokande experiment, which require the neutrino to have a non-zero mass, has rekindled interest in the possibility that neutrinos might, after all, make some contribution to the density of the universe. In order to avoid the problems faced by a universe dominated by hot dark matter outlined above, most of the mass in the universe must still consist of cold dark matter, with neutrinos providing only a minority contribution. Such 'mixed' dark matter models have not proved as successful at matching the data as the pure cold dark matter model. Another possibility for accommodating massive neutrinos is to replace inflation with a different mechanism for generating primordial density fluctuations. Tom Kibble suggested in the 1980s that TOPOLOGICAL DEFECTS IN COSMOLOGY might arise during phase transitions in scalar fields present in the early universe. Certain classes of defects, such as strings and textures, can act as seeds onto which matter gravitates, generating inhomogeneities in the mass density of the universe. These have the defining feature of being non-Gaussian. Unfortunately, modeling the formation of structure in defect models has turned out to be more complicated than in the Gaussian case because, amongst other subtleties, the properties of the defects themselves evolve with time. Defect models have not yet been explored with the same degree of rigor as the cold dark matter model. However, the current indications are that they have difficulties reproducing the spectrum of temperature anisotropies measured in the cosmic microwave background radiation. Thus, from a cosmological point of view, the oscillation data are best accommodated if the neutrino mass is very small (much less than 1 eV), so that neutrinos make a negligible contribution to the cosmic mass budget.

Computer simulations of galaxy formation

The remarkable developments of the past 15 years—the idea of cosmic inflation, the cold dark matter model, the discovery of ripples in the microwave background, and the observations of galaxies at high redshift—have laid down very solid foundations on which to build an understanding of galaxy formation. The 'initial conditions' for the evolution of the dominant dark matter component and its

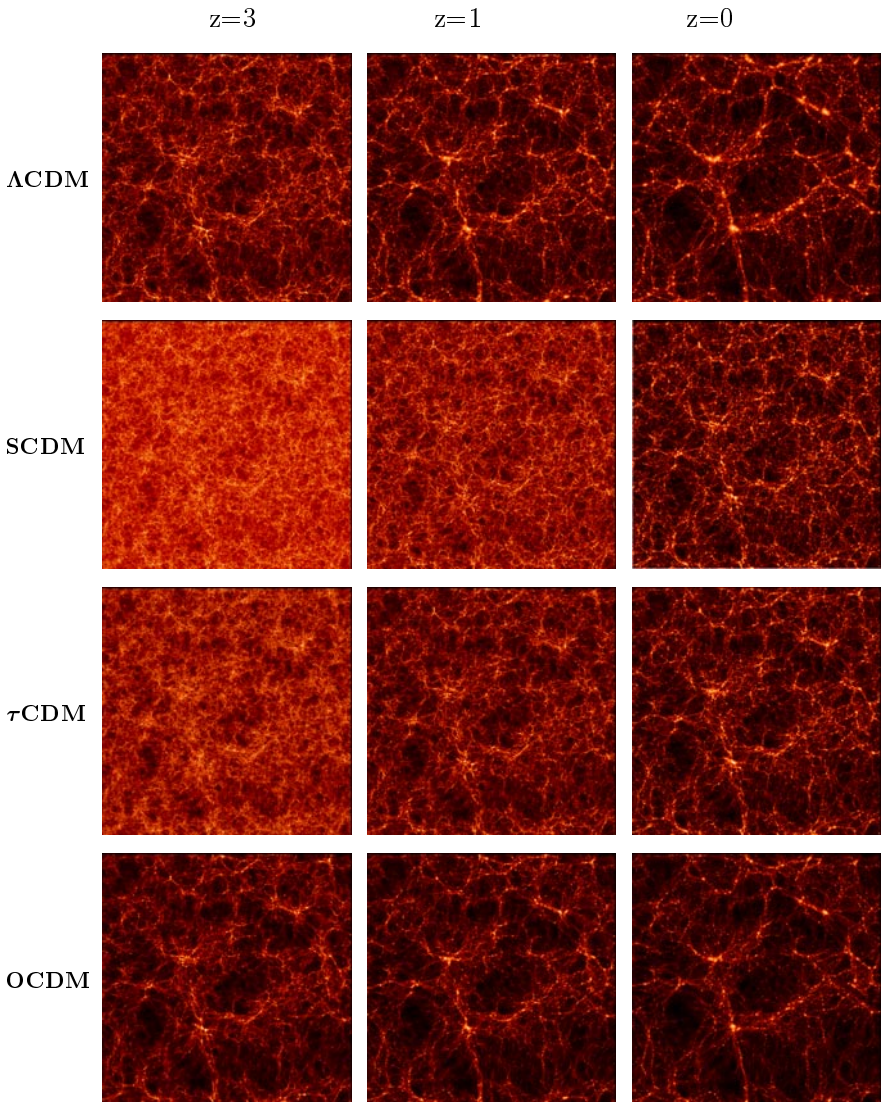

Figure 2. The formation of structure in N-body simulations of representative cosmological volumes of the universe. The intensity of the shading indicates the density of cold dark matter. Each row shows results from different versions of the cold dark matter model. The top row is a flat universe with $\Omega_0 = 0.3$ and a cosmological constant; the middle two rows are $\Omega_0 = 1$ universes, with different power spectra; the bottom row is an open universe with $\Omega_0 = 0.4$. The images, from left to right, show the evolution of structure in each model as a function of redshift. The present day corresponds to $z = 0$ while $z = 3$ corresponds to the epoch when the universe was approximately 15% of its current age. (Courtesy of the VIRGO Consortium for Cosmological N-body Simulations.)

subsequent gravitational evolution are well understood. Yet formulating an *ab initio* theory of galaxy formation and evolution over 10 billion years of cosmic history remains a tall order. The main stumbling block is our poor understanding of the behavior of cosmic gas—most probably a complex, dynamic, multiphase medium, of the physics of star formation, and of the feedback between the two mediated by winds from massive stars and supernovae explosions. The best way to address these issues is through extensive computer simulation and modeling.

The basis for present-day cosmological simulations is the N-body technique, which has been very successfully applied to modeling the evolution of collisionless dark matter. Using various computationally efficient methods, the computer solves the coupled equations of motion of N particles, interacting only through gravity, in the expanding universe. Progress over the past two decades has been driven mainly by dramatic improvements in the speed and memory of computers. By way of illustration, the early simulations of the cold dark matter cosmogony in 1985 employed 32 768 particles. In 1999, the largest simulations performed on massively parallel supercomputers (using essentially the same algorithms as

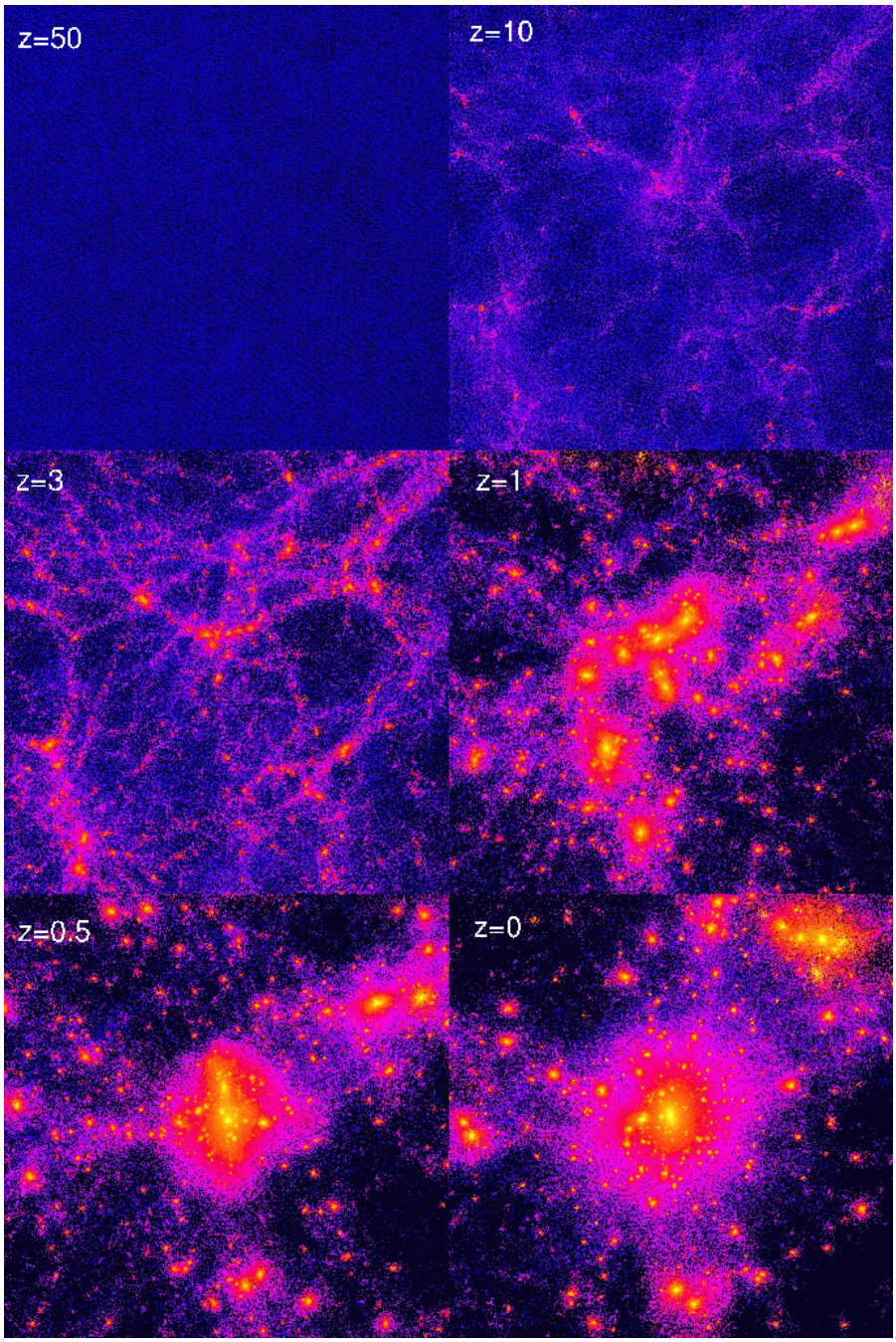

Figure 3. A high-resolution simulation of the formation of a single dark matter halo in a cold dark matter universe. The brighter colors indicate higher densities of dark matter. The sequence shows a series of snapshots of the evolution of the halo, at the redshifts indicated in the legend. The present-day halo displays a significant amount of substructure within its virial radius. (Courtesy of Ben Moore, Joachim Stadel, Tom Quinn and George Lake.) **This figure is reproduced as Color Plate 69.**

those of the 1980s) can follow the evolution of 10^9 particles.

Snapshots from simulations of representative, cosmological volumes are displayed in figure 2. This figure illustrates the evolution of structure in four versions of the cold dark matter model, differing only in the values of the cosmological parameters, Ω and Λ. At the present day, the dark matter is arranged in a complex network of voids, filaments and super-clusters (dubbed the 'Cosmic Web' by Bond, Kofman and Pogosyan). It is similar in all the sim-

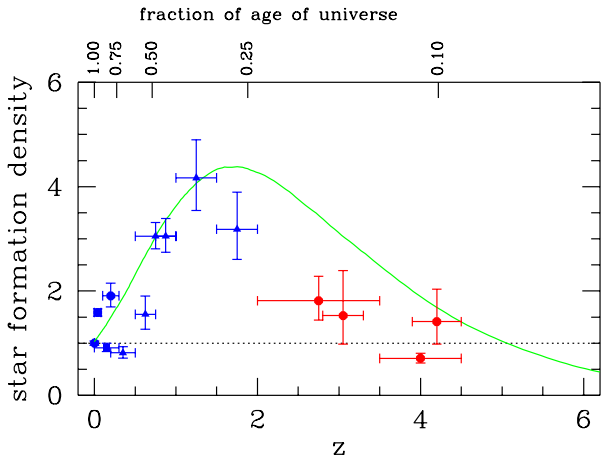

Figure 4. The star formation history of the universe. The curve shows the theoretical prediction for a cold dark matter universe, from the Durham semi-analytic model of galaxy formation. The points show a selection of recent observational determinations of the star formation rate density expressed in units of the present-day value.

ulations, partly because the amplitude of the initial fluctuation spectrum was adjusted so that all models would approximately reproduce the observed local abundance of hot, x-ray emitting clusters. (In addition, the initial fluctuations were set up with the same random phases in all the simulations, so that structures form at the same locations within each volume.) The development of structure proceeds at different rates in the different cosmological models. In those with a low value of Ω (top and bottom rows in figure 2), the formation of structure is essentially frozen at high redshift, whereas in the models with $\Omega = 1$, growth continues to the present. This merely reflects the fact that low-density universes expand more rapidly at late times, with the result that the expansion overwhelms the accretion of matter onto overdensities which lies at the root of their growth.

Simulations of sufficient size to resolve the internal structure and dynamics of dark matter haloes have recently become possible. The largest of these employ several million particles to model the formation of a single halo, revealing the existence of a rich substructure of lumps within the virial radius (figure 3). The simulations show that there is a remarkable uniformity in the density structure of dark matter haloes: over a wide range of scales, the spherically averaged halo density profile in follows a simple form proposed by Navarro, Frenk and White in 1996:

$$\rho(r) \propto \frac{1}{r/r_s(1+[r/r_s]^2)} \qquad (2)$$

where r_s is a scale length related to the density of the universe at the time when the halo formed. This functional form appears to be universal, independent of the choice of halo mass, power spectrum or cosmological parameters.

The N-body technique can be augmented with numerical hydrodynamic methods to model the evolution of gas subject to cooling and heating processes and coupled gravitationally to the dark matter. Two such methods are currently in use: Eulerian methods (including adaptive mesh refinements in some cases) and a Lagrangian scheme known as smoothed particle hydrodynamics. The two techniques have advantages and disadvantages, but so far the latter has proved to be the best suited to studying galaxy formation, because of the formidable dynamic range in thermodynamic quantities involved in the problem. At present, most progress has been achieved in modeling the physics of primitive gas clouds at high redshift, the so-called LYMAN ALPHA FOREST, detected observationally as absorption lines in the spectrum of distant quasars. The first simulations capable of following the evolution of gas to the present, with enough resolution to model the brightest galaxies, are now being carried out by various groups around the world. Currently, only a subset of the relevant gas physics, such as the shock heating of gas within dark matter haloes and its subsequent radiative cooling, are treated reliably.

A complementary technique for simulating galaxy formation *ab initio*, known as semi-analytic modeling, was developed in the 1990s by researchers at Durham and Munich. The main difference with the direct simulation approach is the abandonment of the ideal of solving the equations of hydrodynamics directly, in favor of a simple, spherically symmetric model in which the gas is assumed to have been fully shock-heated to the equilibrium temperature of each halo, from where its cooling and accretion onto the halo can be accurately calculated. This simplification speeds up the calculations enormously and has the added advantage of bypassing resolution considerations which are one of the main limiting factors of full hydrodynamic simulations. Phenomenological models of star formation, feedback and metal enrichment by supernovae are included in the semi-analytic program, through simple scaling relations. The semi-analytic machinery may be grafted onto haloes grown in a cosmological N-body simulation or onto haloes whose formation histories have been generated using a Monte Carlo approach. The models describe the entire star formation and merger history of the galaxy population. The free parameters of the model, which, perhaps surprisingly, are rather few in number, can be set by requiring a good match to a selection of properties of the local galaxy population, such as its luminosity distribution. This results in a fully specified model that provides an ideal tool for comparing the predictions of the cold dark matter theory with observations of the high redshift universe.

Confronting the high-redshift universe

The combination of direct simulations and semi-analytic modeling has revealed in detail the manner in which galaxies are expected to form in the cold dark matter model. The picture that emerges is one of gradual

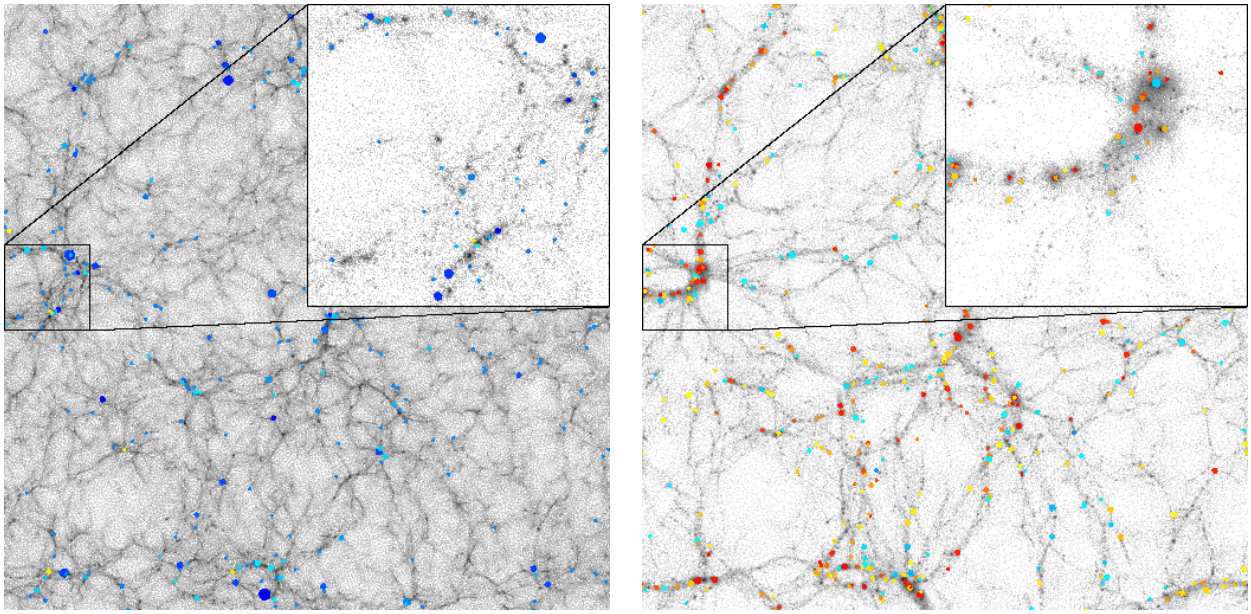

Figure 5. The evolution of clustering in the dark matter and galaxies. The left-hand panel shows an N-body simulation of a flat, low-density, cold dark matter universe with a cosmological constant, in a cube of comoving side $141h^{-1}$ Mpc, at $z = 3$. The right-hand panel shows the same simulation evolved to the present day. The gray scale indicates the density of the dark matter. Dark matter haloes in the simulation have been 'populated' with galaxies using a semi-analytic model of galaxy formation. The color of each spot reflects the color of each model galaxy, which is sensitive to the star formation rate. The size of the spot increases with the absolute luminosity of the galaxy. The inset shows the development of a cluster of galaxies. (Courtesy of Andrew Benson, Shaun Cole, CSF, CMB and Cedric Lacey.) **This figure is reproduced as Color Plates 70 and 71.**

evolution punctuated by major merging events that are accompanied by intense bursts of star formation and which trigger the transformation of disks into spheroids. Galaxy formation stutters into action around $z \sim 5$. Only a tiny fraction of the stars present today would have formed prior to that time. By $z \sim 3$, the epoch when galaxies isolated by the 'Lyman-break' technique[1] are observed, galaxy formation has started in earnest, even though only 10 per cent of the final population of stars has emerged. The midway point is not reached until about a redshift of 1–1.5, when the universe was approximately half of its present age. These theoretical predictions are shown in figure 4. Observationally, the star formation rate per unit volume can be inferred from the density of ultraviolet radiation, which is a measure of the number of high-mass, short-lived stars. Estimates of the star formation density based on data taken by ground-based telescopes and by the Hubble Space Telescope are shown as the points in figure 4. The major uncertainty in the interpretation of these data is the obscuring effect of dust, a modest amount of which has been allowed for in the models. But unless this effect turns out to be much stronger than anticipated, the theory and data in figure 4 suggest that we have now tracked most of the star formation activity, and the associated production of chemical elements, over the entire lifetime of the universe.

A second important prediction of the cold dark matter theory concerns the clustering properties of galaxies at high redshift. At the heart of the hierarchical clustering process lies the fact that galaxies tend to form first near high peaks of the density field because these are the first to collapse at any given epoch. This predilection for high-density regions is known as 'biased galaxy formation' (a term introduced by M Davis in 1985), because the distribution of galaxies offers a biased view of the underlying distribution of mass. An important consequence of biased galaxy formation is that bright galaxies tend to be born in a highly clustered state and remain so for long periods of time. The process of biased galaxy formation is illustrated in figure 5. The left-hand panel shows a snapshot of an N-body simulation of a cold dark matter universe at $z = 3$, whilst the right-hand panel shows the same simulation evolved to the present. The semi-analytic model of galaxy formation has been implemented in the dark matter haloes identified in the simulation at each redshift, in order to populate them with galaxies. Galaxies that are bright enough to be detected at $z = 3$ may be seen to form at the locations where the dark matter density (depicted by the gray scale) is highest. Observational confirmation of this clustering prediction came with the discovery that the population of galaxies at

[1] The so-called 'Lyman-break' galaxies are detected in passbands above the redshifted Lyman-limit at 912 Å and undetected in passbands below this limit; the strength of the Lyman-limit break is enhanced by absorption due to cold gas in the galaxy and in clouds along the line of sight.

$z \sim 3$ identified by the Lyman-break technique is about as strongly clustered as bright galaxies are today. The relative clustering strengths of galaxies and dark matter evolve quite differently. The right-hand panel of figure 5 shows that the dark matter is much clumpier today than it was at $z = 3$. On the other hand, the clustering pattern of galaxies has hardly changed over this long period of time. Galaxies today are found in a wide range of environments and have a clustering amplitude similar to that of the dark matter. This was not the case at high redshift when bright galaxies were much more strongly clustered than the dark matter—in other words, when they were very strongly biased.

The next steps

The two areas of agreement between theory and data highlighted here—the cosmic history of star formation and the clustering of high-redshift galaxies—are particularly noteworthy because they concern fundamental aspects of the theory. The broad agreement between models and data suggests that the main ingredients of a coherent picture of galaxy formation are now in place. These ingredients are: primordial Gaussian density fluctuations; collisionless, non-baryonic dark matter; gravitational instability; and the growth of galaxies by hierarchical clustering. There are justifiably high expectations for the next decade. The number of 10 m class telescopes is proliferating: the first of the four European Space Observatory 'Very Large Telescopes' came into full operation in 1999, the same year in which the Gemini and Subaru telescopes first saw light. Other large telescopes are under construction in the USA and Spain. The middle of the first decade of the new century should also see the launch of NASA's Next Generation Space Telescope, that will search for galaxies out to a redshift of 10, and ESA's Planck Surveyor, that will map the microwave background radiation with unprecedented precision. Towards the end of the decade, the 'Large Millimeter Array', sponsored by a major international partnership, is scheduled to come into operation in the Chilean desert. It will search for galaxy formation at high redshift and examine star formation in nearby galaxies in the still relatively unexplored sub-millimeter waveband. Ultimately, the cornerstone upon which much of modern cosmology rests is the idea that the universe is dominated by non-baryonic dark matter. Experiments under way in the UK, Italy and the USA stand a good chance of detecting it, if it really exists, within the next few years. This will no doubt count as one of the most exciting discoveries in the history of science. For their part, theorists will not be standing still. Increased computing power, more efficient algorithms and, above all, a better understanding of the astrophysics of galaxy formation, are likely to result in a pretty good imitation, by computer, of the processes through which galaxies in our universe formed.

Bibliography

A pedagogical discussion (at an advanced level) of the physics of structure formation in the expanding universe and of the processes that play a role in galaxy formation may be found in:

Peacock J A 1999 *Cosmological Physics* (Cambridge: Cambridge University Press)

Numerical simulations are reviewed in:

Bertschinger E 1998 *Ann. Rev. Astron. Astrophys.* **36** 599–654

Some of the observational data discussed in this article are reviewed in:

Ellis R S 1997 *Ann. Rev. Astron. Astrophys.* **35** 389–443

Carlton M Baugh and Carlos S Frenk