Mon. Not. R. astr. Soc. (1986) **218,** 409–428

Bimodal star formation and remnant-dominated galactic models

Richard B. Larson *Astronomy Department, Yale University, PO Box 6666, New Haven, Connecticut 06511, USA*

Accepted 1985 August 6. Received 1985 August 2; in original form 1985 May 9

Summary. It is proposed that the initial mass function (IMF) for star formation in galaxies is not a monotonic function of stellar mass but is double-peaked or bimodal. In models with a bimodal IMF the star formation rate (SFR) can be strongly decreasing function of time, and stellar remnants can constitute an important or dominant fraction of the total mass. A bimodal model is constructed that accounts for all of the unseen mass in the solar neighbourhood as remnants, and it is shown that this model is consistent with all of the available constraints on the evolution and stellar content of the solar neighbourhood.

The properties of the inner discs of our Galaxy and M83 can be accounted for by similar models in which the high-mass mode of star formation is more dominant than in the solar neighbourhood, and remnants constitute a large fraction of the mass. Bimodal models with a rapidly decreasing SFR and a mass dominated by remnants account better than conventional models for the observed colours, mass-to-light ratios, and gas contents of spiral galaxies, and also allow the increase of both metallicity and mass-to-light ratio with mass in giant elliptical galaxies to be understood. All of the data are consistent with a picture in which the formation of massive stars is favoured at times and in regions where the SFR is high. An extension of such a picture to the earliest stages of star formation in galaxies may allow the dark mass in galactic haloes to be accounted for as remnants of early generations of massive stars.

1 Introduction

A major unsolved problem of galactic astronomy is that approximately half of the mass in the solar neighbourhood has not yet been identified (Bahcall 1984a, b). Because this invisible mass must be in a disc with a scale height not exceeding 700 pc (Bahcall 1984b), it must consist of dissipative matter that settled into a disc and condensed into unseen objects that are either very low-mass stars or stellar remnants. A constraint on the nature of these unseen objects is that their typical mass cannot exceed about $2\,M_{\odot}$, or the wide binaries presently observed in the Galactic disc would have been disrupted (Bahcall, Hut & Tremaine 1985).

Models for the stellar content of the solar neighbourhood have conventionally adopted an initial mass function that is a monotonically declining function of stellar mass, usually approximated by a power law (Salpeter 1955) or by a series of power-law segments (e.g. Tinsley 1980a). It has also often been assumed that the IMF can be extrapolated to small enough masses to allow the unseen mass to be accounted for as very low-mass stars. However, current evidence does not support the existence of a large amount of mass in low-mass stars. This was emphasized by Tinsley (1980b, 1981b), who noted that any reasonable extrapolation of the stellar mass function to lower masses predicts very little additional mass in unseen faint stars. Recent data, as reviewed by Scalo (1986) and Poveda & Allen (1986), indicate that the mass function peaks at a mass near $0.25\,M_\odot$ and drops even more steeply than assumed by Tinsley at masses below $0.2\,M_\odot$, reducing still further the estimated mass in low-mass stars.

The assumption of a monotonic IMF is also challenged by the evidence for a dip in the mass function of nearby stars at a mass of about $0.7\,M_\odot$ (Upgren & Armandroff 1981; Scalo 1986). As was noted by Armandroff (1983), the presence of this dip weakens the 'continuity constraint' that has been used to argue for a nearly constant star formation rate (e.g. Miller & Scalo 1979), and models with a double-peaked IMF and a strongly decreasing star formation rate become possible. In such models, stellar remnants contribute importantly or even dominantly to the total mass. This may be seen from the fact that the model of Tinsley (1981b), which has a constant SFR, predicts a local column density of remnants of $10\,M_\odot\,\mathrm{pc}^{-2}$, which is not negligible compared to the column density of $22\,M_\odot\,\mathrm{pc}^{-2}$ in main-sequence stars. Thus if the past average SFR were three times larger than the present SFR, as in the model favoured by Armandroff (1983), the predicted column density of remnants would be about $30\,M_\odot\,\mathrm{pc}^{-2}$, making them the dominant contributor to the surface density of the Galactic disc. According to Bahcall (1984a), the surface density of unseen matter is about $30\,M_\odot\,\mathrm{pc}^{-2}$; therefore the unseen mass could be entirely accounted for by remnants in a model with a bimodal IMF.

A bimodal IMF has been advocated for other reasons by Güsten & Mezger (1983), who noted that the rate of formation of massive stars in the inner Milky Way as inferred from observations of thermal radio emission is too high to be easily reconciled with models assuming a monotonic IMF. A similar problem was noted for the inner disc of M83 by Talbot (1980) and Jensen, Talbot & Dufour (1981). To resolve this problem, Güsten & Mezger proposed that the IMF is bimodal and has a second peak at a mass of ~ 2–$3\,M_\odot$; if the relative amplitude of this peak is allowed to vary with radius in the Galactic disc, both the surface density of massive stars and the oxygen abundance as a function of radius can be accounted for in a self-consistent way. In such a model, most of the mass in the inner discs of our Galaxy and M83 must be in the form of remnants (see Section 3).

If the IMF in galaxies is bimodal and the SFR declines strongly with time, so that remnants form a large part of the mass, several puzzles relating to stellar populations and the evolution of galaxies can be solved. In the solar neighbourhood, the modest variation of stellar metallicity with age can be explained if the SFR decreases more rapidly for massive stars than for low-mass stars, as in the model of Schmidt (1963); the possibility that star formation initially favoured massive stars was suggested by Schwarzschild & Spitzer (1953). A bimodal model also allows the lifetime of the gas in the solar neighbourhood and in spiral galaxies to be understood without gas infall (Section 2). The colours of the bluest galaxies can be explained if they are dominated by the high-mass mode of star formation, and the weak dependence of mass-to-light ratio on colour can be understood if galactic masses are dominated by remnants (Section 4). Finally, models with a bimodal IMF can account for the increase of both metallicity and mass-to-light ratio with mass in the giant elliptical galaxies (Section 5).

A bimodal IMF could result if low-mass and high-mass stars form in different regions or under different conditions whose relative importance varies with time or location (Mezger & Smith

1977; Larson 1977). Some evidence that this is the case is provided by observations of the nearest regions of star formation, where low-mass stars are seen forming in small cold dark clouds like those in Taurus, while massive stars appear to form only in massive relatively hot molecular clouds like that in Orion. The differences in both the clump mass and the typical stellar mass in Taurus and Orion can be understood at least qualitatively as a result of the strong dependence of the critical mass for fragmentation on the gas temperature (Larson 1985). A possible explanation for bimodality of the IMF based on whether the gas temperature falls or rises at the highest densities is suggested in Section 6. For the present we adopt a bimodal IMF as a working hypothesis, and proceed to consider whether this hypothesis is compatible with all the available constraints. An early suggestion that star formation is bimodal, with low- and high-mass modes that are represented by a 'field star' population and an 'open cluster' population respectively, was made by van den Bergh (1972).

2 Star formation in the solar neighbourhood

2.1 THE LUMINOSITY FUNCTION AND THE IMF

The total amount of mass predicted to be in remnants depends on the total number of stars that have ever formed and died with masses greater than about $1\,M_\odot$, and this can be estimated from the observed luminosity function if an assumption is made about the time dependence of the SFR. If the SFR is assumed to decrease strongly with time, the resulting initial mass function integrated over galactic history is found to increase sharply with mass near $1\,M_\odot$ and to have a second peak at a mass just above $1\,M_\odot$. Schmidt (1959, 1963) and many subsequent authors have argued that such a result is implausible, and that therefore the SFR cannot have been a rapidly decreasing function of time (e.g. Miller & Scalo 1979; Tinsley 1980a). Other authors have accepted a non-monotonic IMF and have suggested that a double-peaked IMF might result from different modes of formation of low-mass and high-mass stars (e.g. Smith, Biermann & Mezger 1978). The latter view becomes more attractive, as we have noted, if the luminosity function has a dip at $0.7\,M_\odot$.

Fig. 1 illustrates the time-integrated IMFs derived from three sets of data using two probably extreme assumptions about the time-dependence of the SFR. The common simplifying assumptions that the IMF is independent of time and that the SFR decays with time as $\exp(-t/\tau)$ have been adopted, and results are shown for $\tau=\infty$ and $\tau=4.15$ Gyr, the latter value corresponding to a past average SFR equal to 10 times the present SFR. Here and throughout the paper, it has been assumed that the age of the Galactic disc is 15 Gyr; this is based on the result of Zinn (1985a) that the most metal-rich globular clusters form a disc system whose age, as judged by the two prototype disc globular clusters M71 and 47 Tuc, is the same as the age of the halo clusters, i.e. at least 15 Gyr. The dots joined by solid lines in Fig. 1 are based on the data compiled by Scalo (1986) for the luminosity function of nearby stars and for the scale heights, masses, and lifetimes of stars as a function of luminosity. Because these data differ quite significantly from those adopted earlier by Miller & Scalo (1979), and because this difference may be an indication of the uncertainties involved, the results derived from the data of Miller & Scalo are also shown, indicated by crosses joined by dashed lines. Finally, the results derived from the data independently compiled by Green (1980) for masses above $0.85\,M_\odot$ are shown for comparison, indicated by plus signs.

It is evident from Fig. 1 that, even for a constant SFR, the time-integrated IMF derived from the more recent data showing the dip at $0.7\,M_\odot$ has a significant bump near $1\,M_\odot$. If the SFR has decreased with time, as is more plausible, the bump is enhanced and the conclusion that the IMF has a second peak just above $1\,M_\odot$ becomes unavoidable. The other extreme case illustrated, which has a past average SFR equal to 10 times the present rate, probably represents too great a

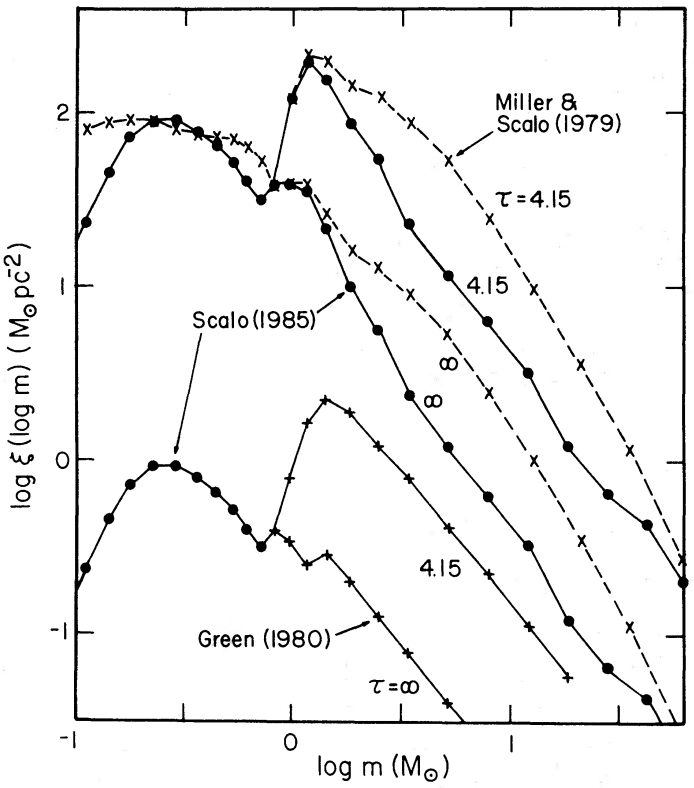

Figure 1. The total number $\xi(\log m)$ of stars per unit log mass ever formed during the past 15 Gyr in a column of cross-section 1 pc^2 through the Galactic plane, as derived from three different sets of data using two assumptions about the time dependence of the SFR. The form of the IMF is assumed not to vary, and the SFR is assumed to decay with time as $\exp(-t/\tau)$, where $\tau=\infty$ or $\tau=4.15$ Gyr, as indicated; the later value of τ corresponds to a ratio of past average to present SFR of 10. Dots, crosses, and plus signs denote data from three different sources, as indicated. The lower set of curves has been shifted downward by two units in $\log\xi$ for clarity.

decrease of the SFR with time because the resulting IMF shows a very steep increase with mass near 1 $M_\odot$. However, an intermediate case with a ratio of past average to present SFR of 3 to 6 could not reasonably be excluded, especially considering the differences between the various sets of data illustrated. An intermediate model somewhere in this range, depending on the assumptions adopted, would provide enough mass in remnants to account for all of the unseen mass in the solar neighbourhood.

Models with a non-monotonic IMF and a large mass in remnants were previously considered by Quirk & Tinsley (1973) and Green (1980), and these authors noted that such models might account for the invisible mass in the solar vicinity. In Section 2.2 we present an illustrative model with a bimodal IMF and a decreasing SFR that accounts for the unseen mass as remnants, and in Sections 2.3–2.5 we consider how this model compares with the available constraints on the star formation history, remnant content, and chemical evolution of the solar neighbourhood.

2.2 A MODEL WITH A BIMODAL IMF

To illustrate how the properties of the solar neighbourhood might be accounted for by a bimodal model, we consider a model in which the IMF consists of two components of similar form representing separate low-mass and high-mass modes of star formation. Because there are significant uncertainties in the fundamental data constraining such models, we follow the spirit of

previous power-law models and assume that each component of the IMF has a simple analytical form, taken to be the same for each mode except for a different characteristic mass. The adopted function approaches a power law at large masses, since a single power law appears to be consistent with all the evidence concerning the IMF of massive stars (Scalo 1986). At low masses an exponential cut-off is assumed that is chosen to match the shape of the peak at $0.25\,M_\odot$ in the time-integrated IMF of Scalo (1986) illustrated in Fig. 1.

The creation function $C(\log m, t)$, defined as the number of stars formed per unit log mass per square parsec per Gyr, is thus assumed to be the sum of two functions C_1 and C_2 representing respectively the low-mass and high-mass modes of star formation:

$$C(\log m, t)=C_1(\log m, t)+C_2(\log m, t). \tag{1}$$

We also assume that the functional dependence of C_1 and C_2 on mass does not vary with time, so that we can write

$$C_k(\log m, t)=B_k(t)F_k(\log m), \tag{2}$$

where $F_k(\log m)$ is normalized to unit total mass; $B_k(t)$ is then the total star formation rate for each mode in $M_\odot\,\mathrm{pc}^{-2}\,\mathrm{Gyr}^{-1}$. The adopted functional form of $F_k(\log m)$ for both the low-mass ($k=1$) and high-mass ($k=2$) modes of star formation is

$$F_k(\log m)=2.55\,m_k m^{-2}\exp\left[-(m_k/m)^{3/2}\right]. \tag{3}$$

The assumed time-dependence of the SFR for each mode is, as before,

$$B_k(t)=A_k\exp\left(-t/\tau_k\right). \tag{4}$$

To fix the parameters of the model, we note first that the rate of formation of stars of nearly solar mass has apparently not varied much with time (see Section 2.3); we have therefore assumed that the SFR of the low-mass mode is constant, i.e. that $\tau_1=\infty$. If only the high mass mode decays with time, the resulting model has the added attraction that it predicts a relatively slow variation of stellar metallicity with age, as is observed (Section 2.5). The remaining parameters for the low-mass mode can then be determined by fitting the time-integrated IMF with $\tau_1=\infty$ to the low-mass peak at $0.25\,M_\odot$ in the mass function illustrated in Fig. 1. This yields $A_1=1.85\,M_\odot\,\mathrm{pc}^{-2}\,\mathrm{Gyr}^{-1}$ and $m_1=0.30\,M_\odot$.

The constraints on the high-mass mode of star formation are that the model should predict enough mass in remnants to account for the unseen mass near the Sun, and it should provide a reasonable fit to the observed stellar mass function for all masses. These requirements do not determine the parameters with complete uniqueness, but a reasonable match to all of the constraints is obtained if $A_2=41\,M_\odot\,\mathrm{pc}^{-2}\,\mathrm{Gyr}^{-1}$, $m_2=2.2\,M_\odot$, and $\tau_2=3.4\,\mathrm{Gyr}$. Assuming a gas column density of $8\,M_\odot\,\mathrm{pc}^{-2}$ and adopting stellar lifetimes from Scalo (1986) and remnant masses from Iben & Renzini (1983), the total column density predicted by the model is then $67\,M_\odot\,\mathrm{pc}^{-2}$, the value favoured by Bahcall (1984a) from dynamical studies. The fit of the model to the observed stellar mass function is illustrated in Fig. 2, where as in Fig. 1 we show the time-integrated IMF

$$\xi(\log m)=\int_0^{15\,\mathrm{Gyr}} C(\log m, t)\,dt. \tag{5}$$

The model prediction is indicated by the heavy curve, and the separate contributions of the low-mass and high-mass modes are shown by the dashed curves. The dots and crosses denote, as in Fig. 1, the values of $\xi(\log m)$ derived from the data of Scalo (1986) and Miller & Scalo (1979), assuming in this case that the time-dependence of the SFR for stars of each mass is the same as predicted by the model. While the data have thus not been 'reduced' in a model-independent

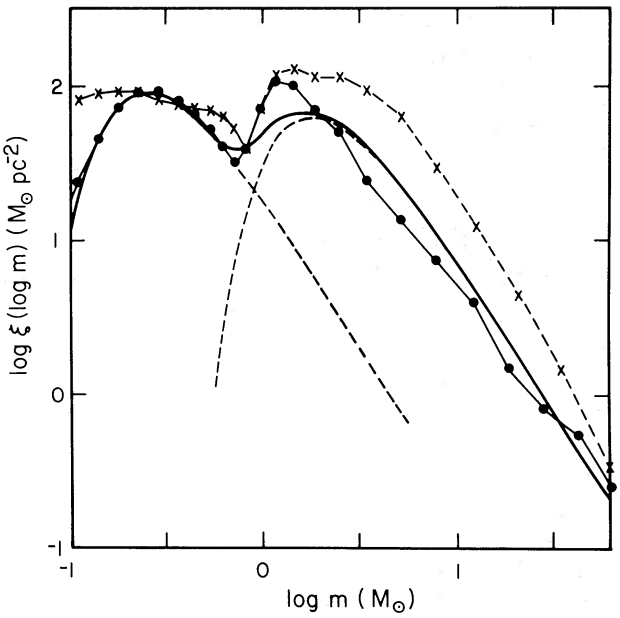

Figure 2. The time-integrated IMF $\xi(\log m)$ predicted by the model of Section 2.2, compared with values of $\xi(\log m)$ derived from two sets of data assuming that the time dependence of the SFR at each mass is the same as predicted by the model. The heavy curve shows the model prediction, and the light dashed curves show the contributions of the low-mass and high-mass modes. The dots represent data from Scalo (1986) and the crosses data from Miller & Scalo (1979). The difference between the heavy curve and the symbols is the same as the difference between the predicted and observed present mass functions.

way, the difference between the model curve and the data points in such a plot is the same as the difference between the predicted and observed present stellar mass functions.

Considering the difference between the two sets of data shown, the agreement between the model and the more recent data of Scalo (1986) in Fig. 2 seems satisfactory. The largest discrepancy occurs near $1\,M_\odot$, where the data yield a more abrupt increase and a sharper peak in $\xi(\log m)$ than the model, but even here the difference does not exceed a factor of 2, which is compatible with the uncertainties in the data; for example, the stellar scale height has a very steep but poorly determined dependence on mass near $1\,M_\odot$. We conclude that a relatively simple bimodal model can account for the local unseen mass while yielding acceptable agreement with the observed luminosity function.

2.3 STAR FORMATION HISTORY AND GAS CONSUMPTION

From a study of the age distribution of nearby F stars, Twarog (1980) concluded that the past average SFR of these stars probably did not exceed twice their present SFR. More indirect arguments for a slowly varying SFR have been given that are based on the metallicity distribution (Twarog 1980) and the velocity distribution (Vader & de Jong 1981) of nearby F and G stars; also, data on stellar lithium abundances and calcium emission-line strengths appear to indicate a slowly varying SFR (Scalo 1986). To compare the model prediction with these constraints, the ratio of the SFR averaged over the past lifetime of stars of each mass to the present SFR has been calculated as a function of stellar mass; this ratio reaches a maximum value of 1.55 for a mass of $0.85\,M_\odot$ and declines rapidly toward 1.00 for smaller or larger masses. Thus, averaged over any finite mass interval, the ratio of past average to present SFR predicted by the model is less than 1.55, consistent with all of the above observational constraints.

The variation with time of the total column density of gas, living stars, and remnants in the

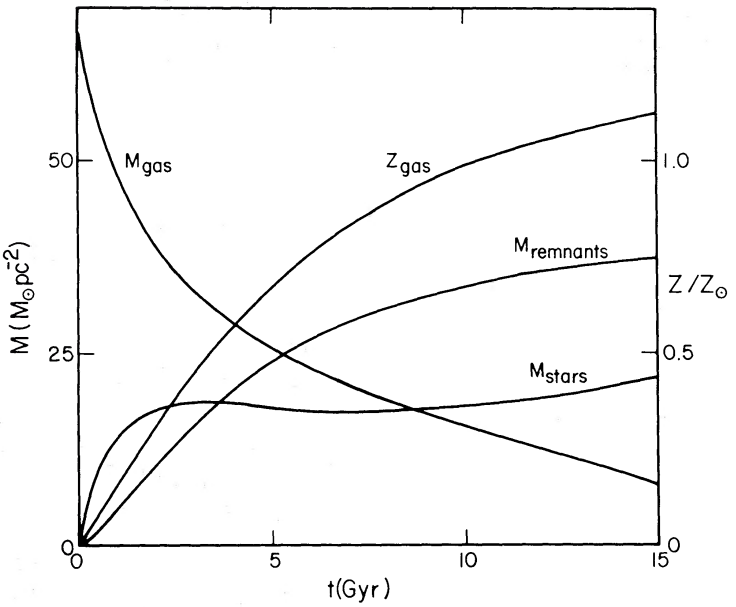

Figure 3. The column density of gas, living stars, and remnants as a function of time for the model of Section 2.2. The metallicity of the gas is also shown as a function of time, normalized to $Z/Z_{\odot}=1$ at $T=10.4$ Gyr. The normalized metallicity has been calculated assuming that the heavy elements are produced only by very massive stars, and that the initial metallicity is zero.

model is shown in Fig. 3. The gas content declines roughly exponentially with time; the total SFR declines somewhat more rapidly, decreasing by a factor of 18 between $t=0$ and $t=15$ Gyr while the gas content decreases by a factor of 8. This corresponds to a decrease of the SFR with about the 1.4 power of the gas content, although a power-law dependence is not closely followed throughout the evolution. The ratio of the past average to the present SFR is 4.7, while for the most massive stars alone this ratio is 13. In these respects, the present model is not greatly different from that of Schmidt (1963), in which the SFR is assumed to be proportional to a power of the gas content that varies from 1 for low-mass stars to ≥ 2 for stars of mass $\geq 10\,M_{\odot}$; the ratio of past average to present SFR is then 2.5 for low-mass stars and 20 for stars of mass $10\,M_{\odot}$. However, the basis of the present model is completely different; no law relating the SFR to the gas content has been assumed (for a critique of such laws see Larson 1977), and the primary constraint on the model is the requirement that it reproduce the amount of dark matter in the solar vicinity as determined dynamically.

Because of the greater importance of gas recycling from massive stars in the present model, and because less mass goes into low-mass stars than in most previous models, the gas depletion rate is smaller than has previously been estimated. This largely removes the problem noted by Larson, Tinsley & Caldwell (1980) and Kennicutt (1983) that the estimated time-scale for gas consumption in the solar neighbourhood and in other galaxies is much shorter than the Hubble time. In the present model, the current time-scale for gas depletion in the solar neighbourhood is $-M_{\rm gas}/\dot{M}_{\rm gas}=5.7$ Gyr; for comparison, a simple exponential decay of the gas content would imply a time-scale for gas depletion of 7.1 Gyr. Since these time-scales are similar, there is no longer a strong argument for gas replenishment by infall, as was suggested by Larson *et al.* (1980).

2.4 REMNANT PROPERTIES

In the model of Section 2.2, living stars presently contribute $22\,M_{\odot}\,{\rm pc}^{-2}$ to the total column density, gas contributes $8\,M_{\odot}\,{\rm pc}^{-2}$, and remnants make up the remaining $37\,M_{\odot}\,{\rm pc}^{-2}$, or 55 per

cent of the total. The predicted amount of mass in remnants is not very sensitive to the assumed relation between remnant mass and initial stellar mass. The above value was calculated using the relation

$$m_{\rm rem}=0.38+0.15\,m_{\rm star} \tag{6}$$

predicted by stellar evolution calculations assuming a standard mass loss rate (Iben & Renzini 1983). With a lower mass loss rate appropriate for Population II stars, the total amount of mass in remnants becomes $42\,M_\odot\,{\rm pc}^{-2}$. If, on the other hand, the remnant mass is assumed to be the larger of $0.6\,M_\odot$ or 15 per cent of the initial stellar mass (Larson 1984), a relation that happens to approximate well the data on white-dwarf masses compiled by Weidemann & Koester (1983), the predicted total mass in remnants becomes $31\,M_\odot\,{\rm pc}^{-2}$. Thus the uncertainty of the total mass in remnants is only about $\pm 6\,M_\odot\,{\rm pc}^{-2}$, which is less than the uncertainty in the dynamical determination of the column density of unseen matter near the Sun.

The mass of a typical remnant in the model is about $0.8\,M_\odot$, and about 4/5 of the total mass in remnants is in white dwarfs. The predicted large number of white dwarfs may conflict with the observed paucity of faint white dwarfs in the solar vicinity (Liebert *et al.* 1979; Liebert, Dahn & Sion 1983); their observed number is too small to be consistent even with conventional models of galactic evolution if one adopts current predictions for the fading of white dwarfs of mass $0.6\,M_\odot$ (Iben & Tutukov 1984). Two possible explanations of this discrepancy are that the fading times of white dwarfs may be shorter than currently estimated, or that the scale height of the oldest white dwarfs may be larger than has been assumed (Iben & Tutukov 1984). If white dwarfs of all masses fade to invisibility in less than 10 Gyr, there is no conflict between the present model and the observations. If this is not the case, it still remains true that white dwarfs more massive than $1\,M_\odot$ should become invisible in less than 10 Gyr (Ostriker & Axel 1969; Lamb & Van Horn 1975). Thus, if there is a conflict it could be removed by assuming that most of the stars formed during early stages of galactic evolution left remnants more massive than $\sim 1\,M_\odot$. A model in which the scale mass m_2 for the high-mass mode of star formation was larger at earlier times is not excluded by any constraints, and is in fact suggested by chemical evolution considerations (Section 2.5).

Shipman (1983) has argued that a large amount of mass cannot be present in invisible remnants because such objects should be detectable through their gravitational effects in binary systems, yet only one case is known of an astrometric binary in which the companion has not been seen (see also Borgman & Lippincott 1983). However, in the present model nearly all of the dark matter results from the high-mass mode of star formation, which produces very few stars of mass $\lesssim 1\,M_\odot$ that would survive as the visible companions of dark objects formed at early times (see Fig. 2). Mass transfer effects would hasten the evolution of many of these low-mass companions, and their number would become even smaller if m_2 were any larger than $2.2\,M_\odot$ at early times. Thus the near-absence of invisible companions in astrometric binaries does not exclude models of galactic evolution like that considered here.

2.5 CHEMICAL EVOLUTION

The time-dependence of the heavy-element abundance predicted by the model has been calculated assuming that the heavy elements are produced only by very massive stars; the result is shown in Fig. 3, normalized to $Z/Z_\odot=1$ at $T=10.4$ Gyr. The predicted $Z(t)$ may be compared with the empirical age–metallicity relation of Twarog (1980) if we note that the assumption of production by massive stars is best justified for oxygen, whereas Twarog's data measure primarily the abundance of iron which may come from less massive stars. The oxygen abundance in disc

stars is observed to vary less than the iron abundance, and we adopt here the relation [O/H]=0.5 [Fe/H] found by Clegg, Lambert & Tomkin (1981). With this conversion between O and Fe abundances, the model agrees well with Twarog's (1980) age–metallicity relation; for example, the predicted increase in [O/H] during the past 10 Gyr is 0.22, compared with the increase of about 0.19 indicated by the observations. Even better agreement is obtained if the initial [O/H] is assumed to be −0.6, as found by Clegg *et al.*; the predicted increase in [O/H] over the past 10 Gyr is then 0.18.

Less good agreement is obtained with the recalibration of Twarog's data proposed by Carlberg *et al.* (1985), which implies an increase in [O/H] over the past 10 Gyr of only about 0.10. A possible explanation of this difference is that the IMF has changed more radically with time than in the present relatively conservative model in which the scale mass m_2 of the high-mass mode of star formation does not vary. If, for example, the value of m_2 at early times were as large as $4.4\,M_\odot$ instead of $2.2\,M_\odot$, the initial oxygen yield would be increased by a factor of 2.6, approximately what would be required to achieve agreement with the revised age–metallicity relation of Carlberg *et al.* (1985). Such a revision of the model, although introducing more parameters, would have the additional advantage of predicting fewer visible white dwarfs (see Section 2.4).

The absolute oxygen abundance has not been used as a constraint on the model because the mass range of stars that explode as supernovae and release oxygen is not well known (e.g. Twarog & Wheeler 1982). The usual assumption that all stars more massive than $10\,M_\odot$ explode (e.g. Arnett 1978) has not yet been confirmed by dynamical calculations, which have yielded definite explosions only for a narrow range of masses near $10\,M_\odot$ (Hillebrandt, Nomoto & Wolff 1984). Since the upper limit of this mass range is uncertain, we have chosen to determine from the model the upper mass limit that would be required if the model is to reproduce the solar oxygen abundance. Using the estimates of Arnett (1978) for the amount of oxygen produced by stars of various masses, we find that only stars less massive than about $16\,M_\odot$ can contribute to oxygen enrichment. The typical mass of an oxygen-producing star is then about $14\,M_\odot$.

From the location of a number of pulsars in or near OB associations, Schild & Maeder (1985) have argued that stars with masses up to $50\,M_\odot$ can explode and leave neutron star remnants, in apparent conflict with the above conclusion that only stars less massive than $16\,M_\odot$ can explode. However, the pre-supernova lifetime of a $16\,M_\odot$ star is only 1.3×10^7 yr, so as long as the associations containing pulsars have been forming stars for at least this length of time, there is no conflict between the above conclusion and the observations.

If stars more massive than $16\,M_\odot$ do not explode and contribute importantly to nucleosynthesis, intermediate-mass stars may play a relatively more important role than in conventional models and may be a major source of elements such as helium, carbon, nitrogen, and iron. In the case of helium, the ratio $\Delta Y/\Delta Z$ of helium enrichment to heavy-element enrichment expected for the present model can be estimated from the theoretical stellar evolution results assembled by Maeder (1983). With a standard IMF, the predicted value of $\Delta Y/\Delta Z$ increases from 1.0 to 2.2 when the upper mass limit for element synthesis is reduced from 150 to $16\,M_\odot$, as in the present model. The value of $\Delta Y/\Delta Z$ will be further increased by helium in winds from massive stars, but this effect should not be large. The resulting (somewhat uncertain) estimate that $\Delta Y/\Delta Z\sim2$–3 for the present model is consistent with observational constraints, and may be in better agreement with them (Serrano & Peimbert 1981) than the value 1.0 predicted for a standard model.

A final aspect of the present model that differs importantly from conventional models is that because of the high rate of gas recycling, deuterium is strongly depleted during the evolution of the system. In the absence of gas replenishment the depletion factor is estimated to be about 40, although this factor could be considerably reduced by even a small amount of infall. It may, in fact, be necessary for deuterium to be depleted by more than an order of magnitude to achieve consistency with predictions of cosmological nucleosynthesis (Audouze 1986).

3 Star formation in the Milky Way and M83

3.1 THE MILKY WAY GALAXY

We next consider the applicability of bimodal remnant-dominated models to other regions in the disc of our Galaxy. The rate of formation of massive stars at each radius in the Galactic disc has been derived from observations of thermal radio emission by Güsten & Mezger (1983), and these data can be used in conjunction with data on the surface density of mass and gas to constrain evolutionary models. An important finding is that the surface density of massive stars increases much more rapidly toward the Galactic centre than the surface density of mass derived from the rotation curve; for example, between the solar radius ($R_\odot$=8 kpc) and R=4 kpc the surface density of massive stars increases by a factor of 15, while the rotation-curve surface density increases by only a factor of 3. Thus the same model cannot apply at all radii, and a radial variation in the IMF is almost certainly required.

A model with a conventional monotonic IMF can be ruled out for the inner Galactic disc because it predicts more mass in low-mass stars than is allowed by the rotation curve. This is shown in the upper panel of Fig. 4, where the curve indicates the surface density derived from the rotation curve by Talbot (1980) and the circles show the surface densities predicted by a model with a monotonic IMF and a constant SFR fitted to the radio data using the assumptions of

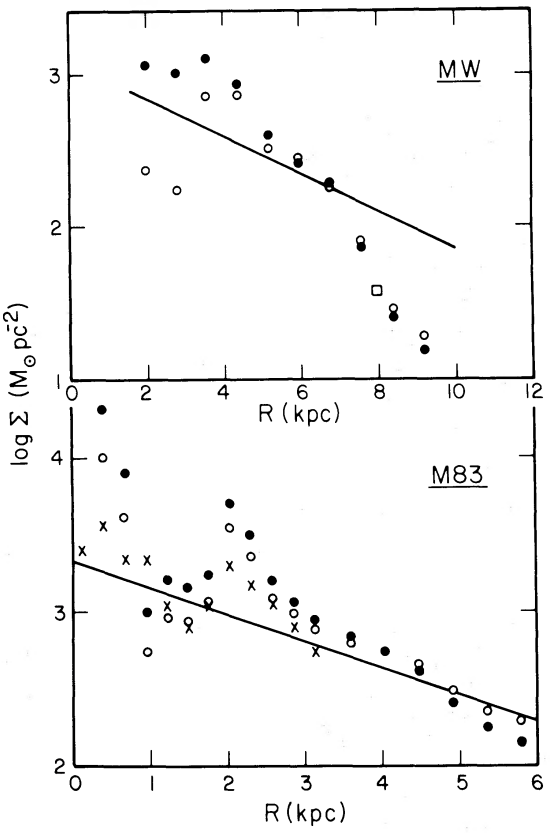

Figure 4. The surface density predicted by various models for the stellar content (symbols), compared with the surface density derived by Talbot (1980) from the rotation curve (solid lines) for both the Milky Way (upper panel) and M83 (lower panel). The circles are for a model with a standard monotonic IMF and a constant SFR fitted to the ionizing photon production rates derived from observations of thermal radio emission (MW) or Hα fluxes (M83). The square in the upper panel is for a similar model fitted to star counts in the solar neighbourhood. The dots are for a model including only a high-mass mode of star formation with m_2=2.2 $M_\odot$ and an SFR that declines in proportion to the gas content. The crosses in the lower panel are for a model with a monotonic IMF and a constant SFR fitted to the *UBV* observations of M83 by Jensen *et al.* (1981).

Güsten & Mezger (1983). The square at $R=8$ kpc shows the surface density given by the similar model of Tinsley (1981b) for the solar neighbourhood, and it is seen that there is approximate agreement, at least locally, between the results derived from the radio data and those based on star counts. The surface density predicted by the model increases strongly with decreasing radius and soon overtakes the rotation-curve surface density, becoming nearly twice as large at $R=4$ kpc. This discrepancy would be even greater if the SFR in the inner Galactic disc has decreased with time, as is likely. Thus the number of low-mass stars that form in the inner Galactic disc must be smaller relative to the number of massive stars than is predicted by a model with a conventional monotonic IMF.

If star formation is more strongly dominated by massive stars in the inner disc, their remnants will contribute an even larger fraction of the total mass than in the solar neighbourhood. Since low-mass star formation is then relatively unimportant, we consider a simplication of the bimodal model of Section 2 in which only the high-mass mode of star formation is retained. Assuming $m_2=2.2\,M_\odot$ as before, the remaining parameter needed to specify a model at each radius is the time-scale τ_2 for decay of the SFR. Since it was found in Section 2 that a simple proportionality of the SFR to the gas content may be approximately valid during the evolution of the solar neighbourhood, we have used this assumption to calculate τ_2 from the gas surface density and the total surface density at each radius.

The dots in the upper panel of Fig. 4 show the surface densities predicted when the above simple model is used with the data of Güsten & Mezger (1983) and the analytic expressions given by these authors for the ionizing photon production rate as a function of stellar mass. The dots follow closely the circles in the outer Galactic disc, but indicate even higher surface densities in the inner disc. Thus a simple model assuming only high-mass star formation with $m_2=2.2\,M_\odot$ predicts *too much mass in remnants alone* in the inner Galactic disc. Possible resolutions of this discrepancy include a ratio of past average to present SFR that is smaller than assumed in the model, or a value of m_2 that is larger than $2.2\,M_\odot$ in the inner Galactic disc. If the latter possibility is correct, the value of m_2 would have to be at least $4.4\,M_\odot$ at $R=4$ kpc in order for the model not to predict too much mass in remnants.

We conclude that, although unique models cannot be determined, remnants almost certainly dominate the mass of the inner Galactic disc; in fact, the condition that there not be *too much* mass in remnants becomes an important constraint on models.

3.2 STAR FORMATION IN M83

Talbot (1980) and Jensen *et al.* (1981) have used measured Hα fluxes to derive the formation rate of massive stars as a function of radius in the nearby spiral galaxy M83, and have noted that the results for M83 are very similar to those for our Galaxy. The lower panel of Fig. 4 shows the surface densities predicted when models similar to those of Section 3.1 are fitted to the Hα fluxes from M83, using again the calibrations of Güsten & Mezger (1983); the curve and the symbols have the same meaning as in the upper panel. The similarity to the inner Milky Way is evident; thus we can conclude, in agreement with Jensen *et al.* (1981), that star formation in the inner disc of M83 must be dominated by stars more massive than a few solar masses. It also follows that most of the mass must be in remnants. In M83, the peaks in the surface density of massive stars at $R=0.4$ and 2.1 kpc are even more pronounced than the peak at $R=4$ kpc in our Galaxy; thus, for example, the model with only high-mass star formation predicts about five times too much mass in remnants alone at $R=2.1$ kpc, indicating that m_2 at this radius could be as large as $8.5\,M_\odot$.

Detailed *UBV* maps of M83 were also obtained by Jensen *et al.* (1981) and were used by them to derive the star formation rate at each radius, assuming a conventional monotonic IMF. The surface densities obtained by multiplying these rates by 15 Gyr are shown by the crosses in the

lower panel of Fig. 4; they may be compared with the circles, which were derived from the Hα fluxes using a nearly identical model. The *UBV* results and the Hα results are seen to be generally consistent, the main difference being that the *UBV* star formation rates show less marked peaks at $R=0.4$ and 2.1 kpc. This difference is qualitatively what would be expected if the peak of the IMF shifts to a higher mass at radii where the local SFR is higher (Larson 1985); this would produce larger changes in the Hα flux, which comes from the most massive stars, than in the *UBV* fluxes, which come from stars nearer the peak in the IMF.

M83 is not exceptional in its high rate of formation of massive stars, since DeGioia-Eastwood *et al.* (1984) have found comparable results for NGC 6946, the main difference being that the SFR in NGC 6946 increases monotonically toward the centre. Further examples of star formation with an IMF that strongly favours massive stars are found in the nuclear regions of some galaxies with high SFRs or 'starbursts' (Rieke *et al.* 1980, 1985); for example, in M82 and NGC 253 the observed star formation activity must be dominated by stars more massive than about $3\,M_\odot$ (Rieke *et al.* 1980; see also Knapp *et al.* 1980). The possibility that the IMF in the blue compact galaxy I Zw 36 lacks stars less massive than about $4\,M_\odot$ was also noted by Viallefond & Thuan (1983), and Augarde & Lequeux (1985) have suggested that recent star formation in Mk 171 has produced only stars more massive than 10 or $20\,M_\odot$. Considering all the evidence, it appears likely that the formation of massive stars is strongly favoured in regions where the local SFR is high. The solar neighbourhood is only a mild example of this phenomenon; the inner discs of our galaxy and M83 are more dominated by massive stars, and the most extreme examples are found in galactic nuclei with particularly high SFRs (including the nucleus of M83; see Fig. 4). The total mass in these regions must, as a result, be dominated by the remnants of massive stars (see also Weedman 1983).

4 Colours and mass-to-light ratios of galaxies

Models of stellar populations that assume a conventional monotonic IMF have been widely used to interpret the colours and mass-to-light ratios of galaxies and to estimate star formation rates (Tinsley 1980a). While broad success has been claimed for these models, there remain several difficulties in accounting for all of the properties of galaxies in this way:

(i) Models with a standard IMF can readily reproduce colours in the range $0.5\lesssim B-V\lesssim 1.0$, but there are many galaxies with bluer colours that cannot be explained unless very young ages are assumed. Some of these blue galaxies are peculiar and probably undergoing bursts of star formation (Larson & Tinsley 1978), but even normal late spiral and irregular galaxies have colours as blue as $B-V\sim 0.25$, and there is no evidence that most of these systems are currently experiencing bursts (Gallagher, Hunter & Tutukov 1984). Similarly, many galaxies have $B-H$ colours considerably bluer than the limiting value of ~ 3.0 predicted by the models of Struck-Marcell & Tinsley (1978); $B-H$ values as small as ~ 1.5 are observed among nearby galaxies (Tully, Mould & Aaronson 1982). These colours cannot be explained by low metallicities alone (Bothun *et al.* 1984), nor does it seem plausible that all of the blue galaxies have very young ages.

(ii) Tinsley (1981a) has noted that the mass-to-light ratios M/L_B of galaxies vary less strongly with $B-V$ than is predicted by standard models, and they are considerably larger than predicted for the bluest galaxies. Tinsley therefore suggested that the proportion of dark matter is highest in the bluest galaxies. Vader (1984) found a similar result in studying the variation of M/L_H with $B-H$; standard models predict that M/L_H decreases with decreasing $B-H$, but the observations show that M/L_H actually increases with decreasing $B-H$.

(iii) Conventional models yield very short time-scales for gas consumption which are difficult

to understand unless gas replenishment by infall is important (Larson *et al.* 1980; Kennicutt 1983). This problem is reduced but not eliminated by the model of Tinsley (1981b), in which the assumption that a substantial amount of gas goes into unseen low-mass stars is dropped; this increases the time-scale for gas consumption by about a factor of 2.

All of the above problems with conventional models can be eliminated if one adopts a model with a bimodal IMF similar to that of Section 2. An IMF that has a second peak at a mass above $1\,M_\odot$ yields a bluer colour than a monotonic IMF for a given star formation history; an IMF enriched in massive stars was suggested, for example, by Sargent & Searle (1970) to explain the colours of blue compact galaxies. The large contribution of remnants to the mass of such a model leads to a higher mass-to-light ratio than a conventional model, which need not depend strongly on colour (see below); the possibility that the contribution of remnants might lead to large mass-to-light ratios for some systems was noted by Freeman (1977). It was already shown in Section 2.3 that a bimodal model alleviates the problem of a short time-scale for gas consumption in the solar neighbourhood, where it is more severe than in most galaxies.

Standard models such as those of Searle, Sargent & Bagnuolo (1973) and Larson & Tinsley (1978) show that the *UBV* colours of galaxies are not sensitive to the details of how the SFR varies with time, and can be adequately reproduced by combining in various proportions an old red population and a blue population with ongoing star formation. To model the colours of galaxies with a bimodal IMF, we therefore adopt a similar two-component model. As a possible blue component, we consider the stars produced by the high-mass mode of star formation of Section 2 proceeding at a constant rate. The colours and the mass-to-light ratio expected for such a population can be estimated from those predicted for a monotonic IMF by Struck-Marcell & Tinsley (1978) by first calculating the luminosity per unit mass in each band for a series of IMFs truncated at various lower mass limits; this can be done by subtracting from the luminosity of each constant-SFR model the luminosity of a single-burst model of the same age containing the same unevolved lower main sequence. The smoothly peaked IMF of the high-mass mode of star formation can then be approximated by adding a series of such truncated IMFs. In this way we estimate that the high-mass population has the colours $B-V=0.20$, $U-B=-0.36$, and $B-H=1.63$, assuming that $H-K=0.2$ for all galaxies (*cf.* Vader 1984). These predicted colours are comparable to those of the bluest normal galaxies (above references; Huchra 1977), so this population is a suitable candidate for the blue population in two-component galaxy models. For the red component, we adopt the colours of a conventional single-burst model (Struck-Marcell & Tinsley 1978), which has $B-V=0.98$, $U-B=0.65$, and $B-H=4.06$.

Although the colours of the blue population were estimated for a constant SFR, they are not very sensitive to the time dependence of the SFR because the blue population consists mostly of relatively short-lived stars. However, the mass-to-light ratio depends on the ratio of past average to present SFR, since most of the mass is in the remnants of evolved massive stars. Since the time dependence of the SFR is not known *a priori*, we regard the mass-to-light ratio of the blue population as a parameter that can be adjusted to fit the observations, and we then deduce from it the time dependence that is required for the SFR. Likewise, the mass-to-light ratio of the red population depends on the unknown number of massive stars formed during the initial burst, so we regard this also as an adjustable parameter.

The colours and mass-to-light ratios of galaxies can be reproduced with a two-component model if the two populations have the following properties:

blue population: $B-V=0.20$, $M/L_B=1.92$, $B-H=1.63$, $M/L_H=2.9$;

red population: $B-V=0.98$, $M/L_B=9.9$, $B-H=4.06$, $M/L_H=1.58$,

where M/L_B and M/L_H are in both cases related by the assumed value of $B-H$. The fit of this

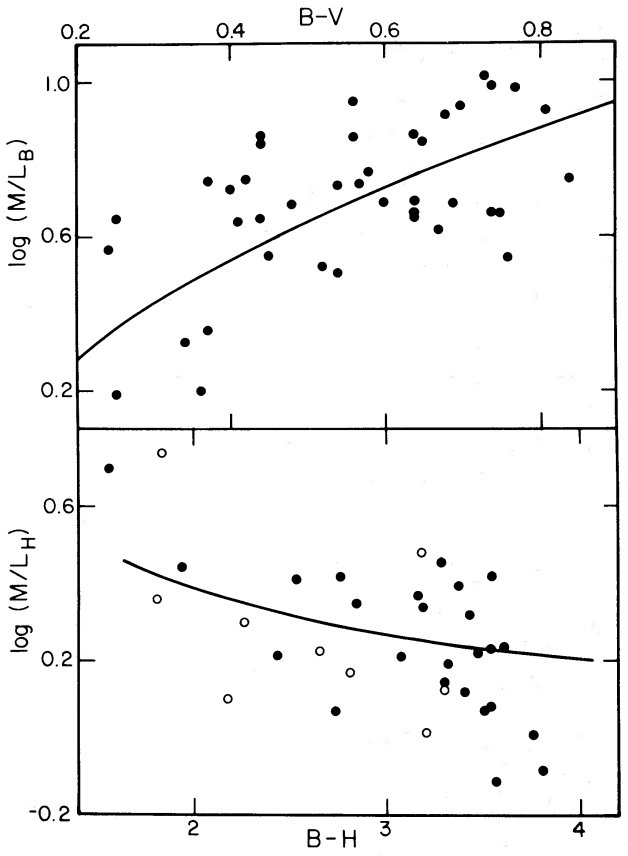

Figure 5. Mass-to-light ratios plotted versus colour for spiral galaxies; the upper panel shows M/L_B versus $B-V$, and the lower panel shows M/L_H versus $B-H$. The points in the upper panel are for the sample studied by Tinsley (1981a), with M/L_B from Faber & Gallagher (1979). The symbols in the lower panel are based on the data assembled by Vader (1984); circles indicate cases in which an extrapolation of the rotation curve was made. The curves in both panels are predicted by the two-component model of Section 4 when the blue and red populations are combined in various proportions.

model to the observations is illustrated in Fig. 5. The upper panel shows M/L_B plotted versus $B-V$ for the same galaxies considered by Tinsley (1981a); the plotted values of M/L_B are from Faber & Gallagher (1979), who assumed a Hubble constant of 50 km s^{-1} Mpc^{-1}. The lower panel shows M/L_H plotted versus $B-H$ for the galaxies considered by Vader (1984); here the values of $B-H$ and L_H have been derived from the measured H magnitudes by using a transformation given by Tully *et al.* (1982) and assuming that $B-H$ is independent of radius. The simple two-component model represents successfully not only the observed range of colours but also the variation of mass-to-light ratio with colour for both visual and infrared wavelengths.

The value of M/L_B predicted for the blue population is 0.32 if the SFR is constant; thus the fact that the observations require that M/L_B for the blue population is ~1.92 means that the SFR must have decreased significantly with time even in the bluest galaxies. If ~60 per cent of the mass of a typical spiral galaxy is in the disc (see below), the bluest disc systems then have $M/L_B \approx 1.15$, which is ~3.6 times the value predicted for a constant SFR; thus the required ratio of past average to present SFR is about 3.6. If the typical gas content of the bluest galaxies is 15 per cent and if a simple exponential decline of the gas content is assumed, the implied ratio of past average to present SFR is 3.0, similar to the value required to account for the mass-to-light ratio. Thus both the colours and the mass-to-light ratios of the bluest galaxies can be plausibly accounted for if they are dominated by a high-mass mode of star formation similar to that of Section 2. For the reddest

systems, the value for M/L_B of 9.9 required to fit the data is about 2.5 times the value predicted by a conventional single-burst model; this difference could plausibly reflect additional mass in the remnants of early generations of massive stars.

According to Malagnini (1979), the integrated $B-V$ colour of a column through the Galactic disc at the solar position is about 0.62; the value of M/L_B for such a column is about 3.5 (Bahcall 1984a, b). From Fig. 5, for typical galaxies with $B-V=0.62$ the value of M/L_B predicted by the two-component model is 5.5, about 1.6 times larger than the solar-neighbourhood value. This difference is similar to the factor of about 1.7 by which the rotation-curve surface density (Fig. 4) exceeds that determined from purely local data, and this suggests that approximately 40 per cent of the mass determining the rotation curve near the Sun lies outside the disc defined by the visible stars. Similar results have been obtained for other spiral galaxies by van der Kruit & Freeman (1986), who estimated mass-to-light ratios for the visible discs of several galaxies and found $M/L_B \sim 4$ (adjusted for $H=50\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$). Comparing this with the total M/L_B values in Fig. 5, we find that typically between one-quarter and one-half of the total mass within the Holmberg radius of a spiral galaxy lies outside the disc defined by the visible stars.

This 'non-disc' mass need not be in a spheroidal distribution, and could just be in a thicker disc than the visible stars. Bahcall (1984b) has noted that the Galactic rotation curve could be fully accounted for if the unseen mass in the solar vicinity were in a disc with a scale height as large as 700 pc. If the earliest star formation occurred in such a thick layer and produced mostly massive stars that left a large amount of mass in remnants, the 'non-disc' matter within the Holmberg radii of spiral galaxies might be explainable by a simple extension of the models discussed here. Some support is given to such a picture by the fact that the scale height of the disc globular clusters is between 500 and 1500 pc (Zinn 1985a); thus these clusters must have formed in a thick disc, and if their formation was accompanied by the formation of large numbers of massive stars, the globular clusters might represent the visible tracers of a population that now consists mostly of unseen remnants.

5 Metallicity variations in galaxies

As was shown by Güsten & Mezger (1983), a model with a bimodal IMF in which the relative amplitude of the two modes varies with radius can account for the radial gradient of the oxygen abundance in our Galaxy. The likelihood that IMF variations are generally responsible for abundance gradients in spiral galaxies is supported by the result of Edmunds & Pagel (1984) that the yield of heavy elements in galactic discs appears to depend only on the local surface density; this suggests that purely local processes of star formation, rather than large-scale gas flows (Lacey & Fall 1985), produce the observed abundance gradients.

Can variability of the IMF also account for the systematic increase of metallicity with mass that is observed among giant elliptical galaxies (Faber 1977)? Although gas loss during early stages of evolution can produce metallicity variations, and is probably required to account for the dwarf ellipticals (see below), gas loss cannot account for the increase of mass-to-light ratio with mass that is also observed in giant elliptical galaxies (e.g. Vader 1986b); this can only be explained by a variation of the IMF. If the increase of both metallicity and mass-to-light ratio with mass is to be explained by a variable IMF, the mass must be more closely associated with the massive stars that produce the heavy elements than with the low-mass stars that provide the light from old populations. This would be the case if most of the mass were in the remnants of massive stars, as in the models of spiral discs discussed earlier. We therefore consider whether bimodal models like that of Section 2 can account for the stellar properties of elliptical galaxies as well.

In a model with a bimodal IMF, both the amount of heavy elements produced and the total mass in stars and remnants are determined largely by the high-mass mode of star formation;

therefore the yield of heavy elements, which depends on the ratio of these quantities, is also determined primarily by the properties of the high-mass mode. If we consider a model with only a high-mass mode of star formation and assume that the IMF has the form given by equation (3), the yield then depends only on the parameter m_2. If we restrict attention again to oxygen and assume a typical mass of $14\,M_\odot$ for oxygen-producing stars (Section 2.5), we calculate that the oxygen yield varies approximately as $m_2^{1.4}$ for values of m_2 near $2.2\,M_\odot$, and it reaches a maximum at $m_2=13\,M_\odot$ that is 5.4 times its value for $m_2=2.2\,M_\odot$. For comparison, the total range in metallicity among giant elliptical galaxies is about a factor of 3 in Fe/H; if the range O/H is also a factor of 3, it could be produced by a variation of a little more than a factor of 2 in m_2. Such a variation in m_2 is plausible in that it is compatible with the likely variations of m_2 with time and location in galaxies that have been discussed in previous sections.

Since the light from an old population is produced by low-mass stars, the mass-to-light ratio depends not only on m_2 but also on the relative amplitudes of the high-mass and low-mass modes of star formation. To estimate how large a variation in relative amplitude is required, we make use of Vader's (1986b) result that metallicity and mass-to-light ratio are correlated with mass such that $M/L_B \propto Z^{1.1}$; if we combine this result with the prediction that $Z \propto m_2^{1.4}$, we obtain $M/L_B \propto m_2^{1.5}$. Since the calculated total mass varies only approximately as $m_2^{0.5}$ for a given amplitude of the high-mass mode, the ratio of amplitudes of the high-mass and low-mass modes must vary approximately as m_2 to account for the observed increase in M/L_B with galactic mass. This requirement is consistent with the evidence discussed previously that the characteristic mass and the relative amplitude of the high-mass mode may vary together in spiral discs. Thus the stellar properties of giant elliptical galaxies can be accounted for with bimodal models very similar to those that have been proposed for spiral galaxies.

Variations of the IMF in a model like that of Section 2 cannot, however, account for the very low metallicities exhibited by the dwarf ellipticals (e.g. Zinn 1985b). The maximum reduction in oxygen yield that can be produced even by the extreme measure of eliminating the high-mass mode of star formation is only about a factor of 20, insufficient to account for the most metal-poor galaxies; in any case, such a change is almost certainly much too extreme because it removes the possibility of explaining the very blue colours of star-forming dwarf galaxies. The very low metallicities of the smallest galaxies therefore probably result from substantial loss of heavy elements during early stages of evolution, plausibly by metal-enhanced winds (Vader 1986a).

6 Summary, discussion, and speculation

In this paper we have proposed a revised view of galactic evolution in which the IMF is bimodal and massive stars formed at a much higher rate in the past, so that stellar remnants now dominate the total mass. The arguments supporting this picture may be summarized as follows:

(i) A model with a bimodal IMF can account for the unseen mass in the solar neighbourhood as remnants. The star formation rate in this model decreases with time roughly as the 1.4 power of the gas content; this is more plausible than the constant or increasing SFR that is implied if the IMF is monotonic.

(ii) If only the SFR of the high-mass mode decreases with time, the model reproduces the age–metallicity relation of Twarog (1980) without requiring additional effects such as gas infall. A similar variation of the IMF with radius can account for the oxygen abundance gradient in our Galaxy, as shown by Güsten & Mezger (1983).

(iii) A bimodal model can account for the high rate of formation of massive stars in the inner regions of our Galaxy and M83 without predicting more mass in low-mass stars than is allowed by the rotation curve. In such a model, most of the mass in these regions is in remnants.

(iv) The colours of the bluest galaxies can be explained if their star formation is dominated by a

high-mass mode. The mass-to-light ratios of galaxies and their weak dependence on colour can also be accounted for if their masses are dominated by remnants.

(v) The present time-scale for gas consumption in a bimodal model is not a small fraction of the Hubble time, as in a conventional model, and is consistent with an exponential decay of the gas content with time.

(vi) The increase of both metallicity and mass-to-light ratio with mass among giant elliptical galaxies can be accounted for by a bimodal model in which the characteristic mass and the relative amplitude of the high-mass mode both increase with increasing galactic mass.

The major implications of this revised scheme of galaxy evolution can be summarized as follows:

(i) Most of the matter that ever goes into stars goes into stars more massive than $1\,M_\odot$. In the model of Section 2, for example, half of the mass goes into stars more massive than $3.6\,M_\odot$. Thus gas recycling is much more important than in conventional models.

(ii) Most of the mass in the visible parts of galaxies is in the remnants of early generations of stars more massive than $1\,M_\odot$. All of the unseen mass within the Holmberg radii of spiral galaxies could be in remnants if the remnants occupy a thicker disc than the visible stars.

(iii) The formation rate of massive stars decreases strongly with time in spiral galaxies, and the ratio of past average to present SFR may exceed 10 for the most massive stars. Thus young spiral galaxies must have been much more luminous than present-day spirals. Forming galaxies must have been exceedingly luminous, probably mostly at infrared wavelengths.

Crucial for this picture is the identity of the unseen objects in the solar vicinity; if they are indeed remnants, the above conclusions follow almost inescapably, whereas if they are low-mass stars, a more conventional picture (and its attendant problems) will remain. Current star formation theory suggests that the unseen mass is more likely to be in remnants, since the minimum critical mass predicted for fragmenting clouds is about $0.3\,M_\odot$ (Larson 1985). A minimum stellar mass that is of this order but decreases with time may even account for the evidence assembled by Poveda & Allen (1986) that stars less massive than $0.2\,M_\odot$ are all relatively young; if the minimum mass for fragmentation decreases with time, as might be expected if the minimum cloud temperature decreases with time, the least massive stars are then also the most recently formed stars.

Can a plausible theoretical basis be given for the two discrete modes of star formation that are required in a bimodal model? As discussed by Larson (1982, 1985), both observations and theory suggest that low-mass stars typically form in small cold dark clouds, while massive stars form in hotter giant molecular clouds. The minimum fragment mass depends sensitively on the thermal behaviour of the gas at the high densities at which the gas temperature couples strongly to that of the dust. The dust temperature is determined by the ambient radiation field, and the dust can be either hotter or colder than the gas depending on whether or not strong sources of radiation are present. In the absence of such sources, the dust is colder than the gas and very low gas temperatures are reached, leading to minimum fragment masses as small as $\sim 0.3\,M_\odot$; this situation could account for the low-mass mode of star formation. If massive stars are present, however, the strong radiation field heats the dust to considerably higher temperatures, causing the gas temperature to rise at the highest densities (Falgarone & Puget 1985) and yielding critical masses for fragmentation that can exceed $10\,M_\odot$ (Larson 1985). Thus a situation in which at least a few massive stars have already formed, perhaps by accumulation processes, could plausibly lead to a high-mass mode of star formation. This mode would be expected to predominate at times and in regions with a high rate of star formation, as would be required to explain the variations of the IMF with time and location that have been discussed.

Finally, we speculate that an extension of the remnant-dominated models considered in this paper could account for the dark matter believed to exist in galactic haloes (Faber & Gallagher 1979). If the earliest stages of star formation in a collapsing protogalaxy produce almost exclusively massive stars that collapse to black holes, and if a substantial fraction of the protogalactic mass is thereby converted into remnants, the resulting galaxy will have a dark halo of remnants surrounding the system of visible stars that form later from the residual gas. A model with a large mass in black holes was first proposed by Truran & Cameron (1971). It is plausible that the earliest stages of star formation should produce almost exclusively massive stars, since both the high rate of star formation and the low metallicity would favour a high gas temperature and hence a large critical mass for fragmentation. Also, the background temperature at early times would have been much higher than at present; at a redshift of 10 it would have been 30 K, high enough to raise the critical mass to values above $100\,M_\odot$. Thus stars formed at such times would almost certainly have been very massive, and would have collapsed to black holes.

An attractive feature of such a picture, in which the visible disc of a spiral galaxy may represent only a small fraction of the initial protogalactic mass, is that the galaxy collapse models of Larson (1975, 1976) always yielded systems in which only a small fraction of the mass was a disc component unless large *ad hoc* modifications were made in the assumed star formation rate. If most of the galactic mass is actually in a dark halo of remnants, these collapse models may apply without such large *ad hoc* modifications, the only revision required being that the IMF varies with time such that initially only massive stars form, while significant formation of low-mass stars does not occur until the residual gas has settled into a disc. Such a disc would then be a secondary system formed partly of matter recycled from halo stars, as proposed by Ostriker & Thuan (1975). The possibility that the dark matter and the visible matter in galaxies are closely related has also been suggested on dynamical grounds by Bahcall & Casertano (1985).

We conclude that the basic properties of the stellar populations in galaxies can be understood in a relatively simple way if the IMF is bimodal and galactic masses are dominated by the unseen remnants of early generations of massive stars. The visible stars then form just the 'tip of the iceberg'.

Acknowledgments

I am indebted to J. P. Vader, R. J. Zinn, and J. P. Ostriker for discussions and information, and especially to my late colleague B. M. Tinsley, on whose pioneering contributions much of the present work is based.

References

Armandroff, T. E., 1983. *The Nearby Stars and the Stellar Luminosity Function, IAU Colloq. No. 76,* p. 229, eds Philip, A. G. D. & Upgren, A. R., L. Davis Press, Schenectady.
Arnett, W. D., 1978. *Astrophys. J.,* **219,** 1008.
Audouze, J., 1986. *Dark Matter in the Universe, IAU Symp. No. 117,* Reidel, Dordrecht, Holland, in press.
Augarde, R. & Lequeux, J., 1985. *Astr. Astrophys.,* **147,** 273.
Bahcall, J. N., 1984a. *Astrophys. J.,* **276,** 169.
Bahcall, J. N., 1984b. *Astrophys. J.,* **287,** 926.
Bahcall, J. N. & Casertano, S., 1985. *Astrophys. J.,* **293,** L7.
Bahcall, J. N., Hut, P. & Tremaine, S., 1985. *Astrophys. J.,* **290,** 15.
Borgman, E. R. & Lippincott, S. L., 1983. *Astr. J.,* **88,** 120.
Bothun, G. D., Romanishin, W., Strom, S. E. & Strom, K. M., 1984. *Astr. J.,* **89,** 1300.
Carlberg, R. G., Dawson, P. C., Hsu, T. & VandenBerg, D. A., 1985. *Astrophys. J.,* **294,** 674.
Clegg, R. E. S., Lambert, D. L. & Tomkin, J., 1981. *Astrophys. J.,* **250,** 262.
DeGioia-Eastwood, K., Grasdalen, G. L., Strom, S. E. & Strom, K. M., 1984. *Astrophys. J.,* **278,** 564.

Edmunds, M. G. & Pagel, B. E. J., 1984. *Mon. Not. R. astr. Soc.,* **211,** 507.
Faber, S. M., 1977. *The Evolution of Galaxies and Stellar Populations,* p. 157, eds Tinsley, B. M. & Larson, R. B., Yale University Observatory, New Haven.
Faber, S. M. & Gallagher, J. S., 1979. *Ann. Rev. Astr. Astrophys.,* **17,** 135.
Falgarone, E. & Puget, J. L., 1985. *Astr. Astrophys.,* **142,** 157.
Freeman, K. C., 1977. *The Evolution of Galaxies and Stellar Populations,* p. 133, eds Tinsley, B. M. & Larson, R. B., Yale University Observatory, New Haven.
Gallagher, J. S., Hunter, D. A. & Tutukov, A. V., 1984. *Astrophys. J.,* **284,** 544.
Green, R. F., 1980. *Astrophys. J.,* **238,** 685.
Güsten, R. & Mezger, P. G., 1983. *Vistas Astr.,* **26,** 159.
Hillebrandt, W., Nomoto, K. & Wolff, R. G., 1984. *Astr. Astrophys.,* **133,** 175.
Huchra, J. P., 1977. *Astrophys. J.,* **217,** 928.
Iben, I. & Renzini, A., 1983. *Ann. Rev. Astr. Astrophys.,* **21,** 271.
Iben, I. & Tutukov, A. V., 1984. *Astrophys. J.,* **282,** 615.
Jensen, E. B., Talbot, R. J. & Dufour, R. J., 1981. *Astrophys. J.,* **243,** 716.
Kennicutt, R. C., 1983. *Astrophys. J.,* **272,** 54.
Knapp, G. R., Phillips, T. G., Huggins, P. J., Leighton, R. B. & Wannier, P. G., 1980. *Astrophys. J.,* **240,** 60.
Lacey, C. G. & Fall, S. M., 1985. *Astrophys. J.,* **290,** 154.
Lamb, D. Q. & Van Horn, H. M., 1975. *Astrophys. J.,* **200,** 306.
Larson, R. B., 1975. *Mon. Not. R. astr. Soc.,* **173,** 671.
Larson, R. B., 1976. *Mon. Not. R. astr. Soc.,* **176,** 31.
Larson, R. B., 1977. *The Evolution of Galaxies and Stellar Populations,* p. 97, eds Tinsley, B. M. & Larson, R. B., Yale University Observatory, New Haven.
Larson, R. B., 1982. *Mon. Not. R. astr. Soc.,* **200,** 159.
Larson, R. B., 1984. *Mon. Not. R. astr. Soc.,* **210,** 763.
Larson, R. B., 1985. *Mon. Not. R. astr. Soc.,* **214,** 379.
Larson, R. B. & Tinsley, B. M., 1978. *Astrophys. J.,* **219,** 46.
Larson, R. B., Tinsley, B. M. & Caldwell, C. N., 1980. *Astrophys. J.,* **237,** 692.
Liebert, J., Dahn, C. C., Gresham, M. & Strittmatter, P. A., 1979. *Astrophys. J.,* **233,** 226.
Liebert, J. W., Dahn, C. C. & Sion, E. M., 1983. *The Nearby Stars and the Stellar Luminosity Function, IAU Colloq. No. 76,* p. 103, eds Philip, A. G. D. & Upgren, A. R., L. Davis Press, Schenectady.
Maeder, A., 1983. *ESO Workshop on Primordial Helium,* p. 89, eds Shaver, P. A., Kunth, D. & Kjär, K., European Southern Observatory, Munich.
Malagnini, M. L., 1979. *Astrophys. Space Sci.,* **60,** 305.
Mezger, P. G. & Smith, L. F., 1977. *Star Formation, IAU Symp. No. 75,* p. 133, eds de Jong, T. & Maeder, A., Reidel, Dordrecht, Holland.
Miller, G. E. & Scalo, J. M., 1979. *Astrophys. J. Suppl.,* **41,** 513.
Ostriker, J. P. & Axel, L., 1969. *Low Luminosity Stars,* p. 357, ed. Kumar, S. S., Gordon & Breach, New York.
Ostriker, J. P. & Thuan, T. X., 1975. *Astrophys. J.,* **202,** 353.
Poveda, A. & Allen, C., 1986. Preprint.
Quirk, W. J. & Tinsley, B. M., 1973. *Astrophys. J.,* **179,** 69.
Rieke, G. H., Lebofsky, M. J., Thompson, R. I., Low, F. J. & Tokunaga, A. T., 1980. *Astrophys. J.,* **238,** 24.
Rieke, G. H., Cutri, R. M., Black, J. H., Kailey, W. F., McAlary, C. W., Lebofsky, M. J. & Elston, R., 1985. *Astrophys. J.,* **290,** 116.
Salpeter, E. E., 1955. *Astrophys. J.,* **121,** 161.
Sargent, W. L. W. & Searle, L., 1970. *Astrophys. J.,* **162,** L155.
Scalo, J. M., 1986. *Fundam. Cosmic Phys.,* in press.
Schild, H. & Maeder, A., 1985. *Astr. Astrophys.,* **143,** L7.
Schmidt, M., 1959. *Astrophys. J.,* **129,** 243.
Schmidt, M., 1963. *Astrophys. J.,* **137,** 758.
Schwarzschild, M. & Spitzer, L., 1953. *Observatory,* **73,** 77.
Searle, L., Sargent, W. L. W. & Bagnuolo, W. G., 1973. *Astrophys. J.,* **179,** 427.
Serrano, A. & Peimbert, M., 1981. *Rev. Mexicana Astr. Astrof.,* **5,** 109.
Shipman, H. L., 1983. *The Nearby Stars and the Stellar Luminosity Function, IAU Colloq. No. 76,* p. 417, eds Philip, A. G. D. & Upgren, A. R., L. Davis Press, Schenectady.
Smith, L. F., Biermann, P. & Mezger, P. G., 1978. *Astr. Astrophys.,* **66,** 65.
Struck-Marcell, C. & Tinsley, B. M., 1978. *Astrophys. J.,* **221,** 562.
Talbot, R. J., 1980. *Astrophys. J.,* **235,** 821.

Tinsley, B. M., 1980a. *Fundam. Cosmic Phys.*, **5,** 287.
Tinsley, B. M., 1980b. *Astr. Astrophys.*, **89,** 246.
Tinsley, B. M., 1981a. *Mon. Not. R. astr. Soc.*, **194,** 63.
Tinsley, B. M., 1981b. *Astrophys. J.*, **250,** 758.
Truran, J. W. & Cameron, A. G. W., 1971. *Astrophys. Space Sci.*, **14,** 179.
Tully, R. B., Mould, J. R. & Aaronson, M., 1982. *Astrophys. J.*, **257,** 527.
Twarog, B. A., 1980. *Astrophys. J.*, **242,** 242.
Twarog, B. A. & Wheeler, J. C., 1982. *Astrophys. J.*, **261,** 636.
Upgren, A. R. & Armandroff, T. E., 1981. *Astr. J.*, **86,** 1898.
Vader, J. P., 1984. *Formation and Evolution of Galaxies and Large Structures in the Universe,* Third Moriond Astrophysics Meeting, p. 227, eds Audouze, J. & Van, J. Tran Thanh, Reidel, Dordrecht, Holland.
Vader, J. P., 1986a. *Astrophys. J.*, in press.
Vader, J. P., 1986b. *Astrophys. J.*, in press.
Vader, J. P. & de Jong, T., 1981. *Astr. Astrophys.*, **100,** 124.
van den Bergh, S., 1972. *External Galaxies and Quasi-Stellar Objects, IAU Symp. No. 44,* p. 1, ed. Evans, D. S., Reidel, Dordrecht, Holland.
van der Kruit, P. C. & Freeman, K. C., 1986. *Astrophys. J.*, in press.
Viallefond, F. & Thuan, T. X., 1983. *Astrophys. J.*, **269,** 444.
Weedman, D. W., 1983. *Astrophys. J.*, **266,** 479.
Weidemann, V. & Koester, D., 1983. *Astr. Astrophys.*, **121,** 77.
Zinn, R. J., 1985a. *Astrophys. J.*, **293,** 424.
Zinn, R. J., 1985b. *Mem. Soc. astr. Ital.*, **56,** 223.

Note added in proof

F. N. Bash & H. C. D. Visser (*Astrophys. J.*, **247,** 488, 1981) found in fitting density-wave models to the detailed photometry of M81 by F. Schweizer (*Astrophys. J. Suppl.*, **31,** 313, 1976) that the spiral-arm colours can be accounted for only if the IMF in the arms is strongly enhanced in massive stars. This result is consistent with the data on our Galaxy and M83 discussed in this paper, and with the suggestion (Mezger & Smith 1977; Larson 1977) that star formation in spiral arms strongly favours massive stars.