conventional computational fluid dynamics are still in the developmental stage in ocean modeling; their efficacy is still largely unproven.

The finite-difference grid can be staggered or nonstaggered, with the former being more accurate. There are several possibilities [Mesinger and Arakawa 1976], including the so-called Arakawa C-grid, where the velocity component U_1 is displaced half a grid to the west and U_2 half a grid to the south of the grid center where all scalar quantities such as η, Θ, and S reside. The C-grid has better wave propagation characteristics if the grid size is smaller than the Rossby radius of deformation. The Arakawa B-grid, where both velocities are displaced half a grid point to the west, is better if it is larger [Semtner 1986]. With increased computing power and hence finer resolution, C-grid is becoming more popular.

Explicit or implicit methods can be used for time-stepping the equations; the latter are more efficient but require more complex simultaneous solution at all model grid points. The former are more easily adapted to massively parallel processors and are being increasingly used despite the limitation imposed by numerical stability considerations. The maximum time step that can be taken in an explicit scheme (for a staggered grid) is given by the Courant–Friedrichs–Lewy (CFL) condition, which is of the form

$$\Delta t \leq 0.5(\Delta x_e / C_e) \tag{31.15}$$

where

$$\Delta x_e = \left(1/(\Delta x_1)^2 + 1/(\Delta x_2)^2\right)^{-1/2}$$

is the effective grid size, which is smaller than the grid size in the individual directions, and C_e is the effective gravity-wave speed, which is the sum of the gravity-wave speed and the advection velocity. In the barotropic problem, for example, $C_e = \max[|U_j| + \sqrt{gH}]$.

Explicit inclusion of the free-surface dynamics in a model requires that a mode-splitting technique [Blumberg and Mellor 1987, Kantha and Piacsek 1993, Madala and Piacsek 1977] be employed to overcome the severe limitations on the solution due to stability considerations imposed by fast-moving external gravity waves on the free surface. This technique consists essentially of splitting the solution into barotropic and baroclinic modes, with the barotropic part solved at the time step dictated by external gravity waves and the baroclinic part at a much larger time step, 20 to 50 times larger. This approach takes into account the fact that internal baroclinic adjustments are much slower.

It is the discretization of the vertical coordinate that is the most distinguishing feature of various ocean models. Several choices are possible, including that of no discretization (for a barotropic model). We will describe these next.

31.3.1 Barotropic Models

If the density gradients are neglected in the governing equations, or alternatively the ocean is considered to be of uniform density, the current distribution in the vertical becomes independent of depth (away from regions of frictional influence such as the surface and the bottom). Under these conditions, it is possible to ignore the transport equations for Θ and S and integrate the governing equations for continuity and momentum over the water column to arrive at a vertically integrated set of equations that govern the sea-surface elevation η and the vertically averaged velocity components $\overline{U}_j$:

$$\begin{aligned}
\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x_k}(\overline{U}_k D) &= 0 \\
\frac{\partial}{\partial t}(\overline{U}_j D) + \frac{\partial}{\partial x_k}(\overline{U}_j \overline{U}_k D) + f\varepsilon_{j3k}(\overline{U}_k D) &= -gD\frac{\partial \eta}{\partial x_j} - D\frac{\partial P_a}{\partial x_j} \\
&\quad + gD\frac{\partial \xi}{\partial x_j} + (\tau_{0j} - \tau_{bj}) + D\overline{F}_j
\end{aligned} \tag{31.16}$$

where $D = H + \eta$ is the total depth of the water column.

The bottom friction is now determined using the column average velocity $\overline{U}_j$. Note the presence of tidal potential terms involving ξ on the right-hand side of the momentum equations that contain astronomical forcing terms due to the gravitational forces of the moon and the sun. Note also the terms due to atmospheric pressure and wind stress forcing. The astronomical forcing can be prescribed a priori from a knowledge of the ephemerides of the sun and the moon [see, for example, Kantha 1995, Schwiderski 1980]. The atmospheric forcing terms are also known and can be prescribed as a function of time during the model run. This set of equations can be used to solve for the sea surface height (SSH) and depth-averaged currents due to phenomena such as tides and storm surges.

Figure 31.1 shows an example of the application of barotropic equations to the problem of deducing the tidal SSH in the global oceans. The reader is referred to Kantha [1995] for details, but, briefly, the equations are cast in spherical coordinates, and the tidal potential terms are expressed as a sum of a series containing various tidal components such as the semidiurnal M_2, with a period of 12.42 h, and the diurnal K_1, with a period of 23.93 h (the atmospheric forcing terms are zeroed out for this application). The resulting equations are solved on a $\frac{1}{5}^\circ$ latitude–longitude C-grid covering the global oceans (excluding the Arctic) for each tidal component. The bottom depths over the model grid are derived from a digital database (ETOP05 from NOAA) containing world topography at $\frac{1}{12}^\circ$ resolution. However, for the results to be accurate enough for certain applications such as altimetry, inevitable errors that result from inaccurate knowledge of bottom depths and friction coefficients have to be overcome by data assimilation. Tidal SSHs can be derived in the deeper parts of the oceans quite accurately from measurements of SSH fluctuations by a satellite-borne microwave altimeter. The tidal SSH data derived from the currently operational NASA/CNES TOPEX/Poseidon precision **altimeter** [Desai and Wahr 1995] have been assimilated into the model as well as those from coastal tide gauges around the world's coastlines. A simple data assimilation scheme has been used where, at each time step, the model-predicted SSH is replaced by a weighted sum of the model SSH and the observed SSH, with weights determined a priori. The result is tidal SSH that is accurate to within a few centimeters over the global oceans, including shallow coastal and semienclosed seas. This information is useful for many applications, such as an accurate determination of the subtidal SSH variability from altimetric data, gravimetry, and determination of tidal dissipation. Figure 31.1 shows the M_2 **coamplitude and cophase** (with respect to Greenwich) distributions of the tidal SSH and the tidal-current ellipses over the global oceans. Figure 31.2 shows the accuracy attained by this data-assimilative tidal model in the form of scatterplots of comparison of modeled and observed tides from an independent set of accurate tide and bottom-pressure gauges over the global oceans, whose locations are also shown.

Barotropic models such as these can also be used to study the response of the SSH to atmospheric pressure forcing [Kantha et al. 1994, Ponte 1994]. It is often assumed that the ocean responds instantaneously to pressure forcing as an inverse barometer with roughly one centimeter of increase (decrease) for every millibar of drop (rise) in atmospheric pressure. This is not always true, and the departures from the inverse-barometer response are quite important to satellite ocean altimetry [Kantha et al. 1994].

Finally, a very important application of barotropic models is for prediction of storm surge effects along a coastline due to approaching hurricanes. The strong hurricane-force winds (augmented by the pressure drop in the eye of the hurricane) pile up water against the coast that often leads to an increase in sea level of several meters and consequent inundation of structures along the coastline. Hurricane Camille in 1969 caused a storm surge of nearly 8 m along the Mississippi coast, leading to widespread destruction and devastation. Provided the local bathymetry is known accurately and the characteristics of the hurricane (such as the wind stress distribution and forward velocity) can be deduced reasonably well from NWP forecasts, it is possible to predict the resulting storm surge quite accurately using a barotropic model driven by the wind stress and atmospheric pressure terms on the right-hand side.

31.3.2 *z*-Level Models

The Bryan–Cox–Semtner z-level model [Bryan 1969, Cox 1985, Semtner and Chervin 1992] is the oldest and the most popular global ocean model. Several versions exist, including the Modular Ocean Model (MOM) from the Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey, the latest version

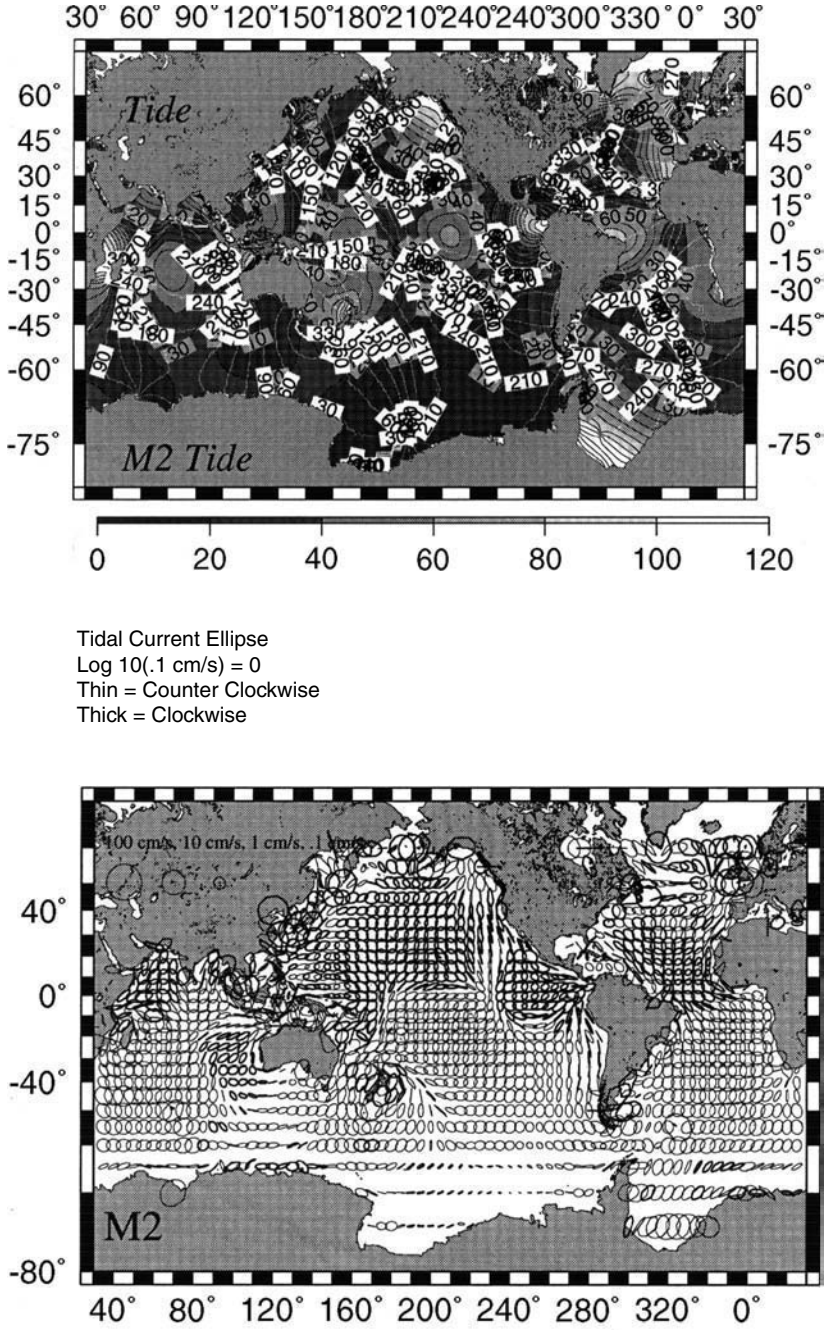


FIGURE 31.1 A map of the distribution of coamplitude and cophase (top), and tidal-current ellipses plotted every 25th point in each direction (bottom) for the M_2 tidal component in the global oceans. Note the logarithmic scale for ellipses.

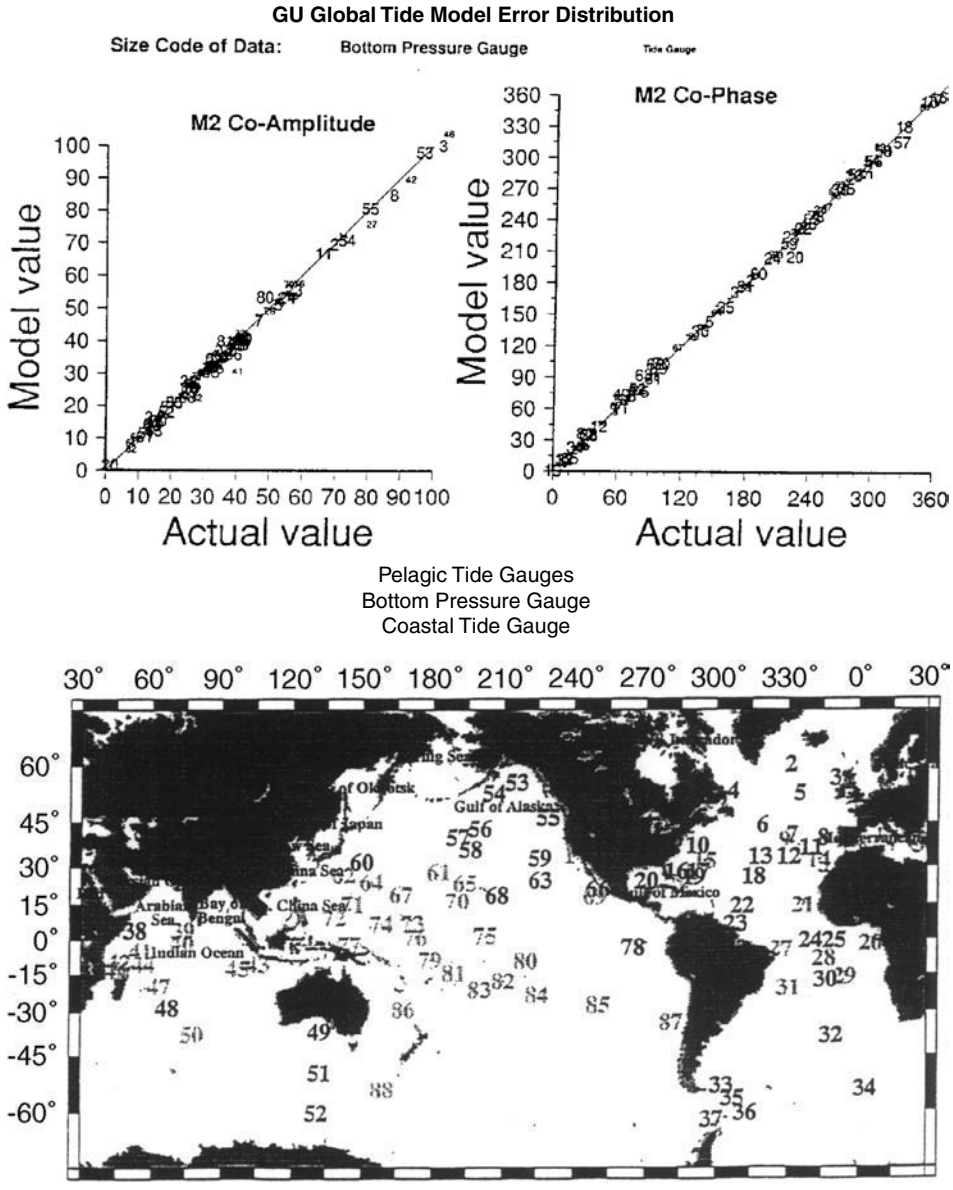


FIGURE 31.2 Scatterplots (top) of modeled M_2 coamplitudes and cophases vs. those observed at pelagic tidal stations, the locations of which are shown at the bottom (darker numbers: bottom pressure gauges; lighter ones: coastal tide gauges).

of which (MOM2) includes free-surface dynamics. A version optimized for massively parallel processors, called POPS (Parallel Ocean Prediction System), is available from Los Alamos Scientific Laboratory [Smith et al. 1992]. Recent improvements include inclusion of a free surface in a variety of versions around the world [Killworth et al. 1991, Dukowicz and Smith 1994], and adoption of a C-grid. A recent review of the current state of ocean modeling using z-level models can be found in Semtner [1995].

Imposition of a rigid lid (via the boundary condition $W = 0$ at $z = 0$) means that the pressure at $z = 0$ enters as an unknown. It is eliminated from Equation 31.2 by cross-differentiation, and an equation for the stream function ψ for vertically integrated transport in the water column is derived (for details, see

Semtner [1986]). This is an elliptic equation and is solved subject to conditions imposed on lateral ocean boundaries, which are in general multiply connected. Herein lies the principal problem with rigid-lid models. While they are efficient, the solution technique is more complicated and not easily adapted to vector and parallel processors. The problem is alleviated somewhat by not cross-differentiating to derive the stream function, but working with the pressure on the rigid lid (Section 31.2, "Rigid Lid or Free Surface").

Numerous applications of the z-level Bryan–Cox–Semtner model and its various versions can be found in the literature (for example in the *Journal of Physical Oceanography* and the *Journal of Geophysical Research, Oceans*). It has been used extensively to study the seasonal, interannual, and climatic variations in the global oceans. It is also a central part of the ocean analysis system for the tropical oceans [Leetma and Ji 1989], where a best estimate of the state of these oceans is determined by assimilation of observational data into a tropical-ocean version of the model. The most recent application can be found in [Semtner and Chervin 1992]. The highest-resolution global z-level model at present is the $\frac{1}{6}^{\circ}$ POPS model at Los Alamos that is run on a 256-node CM5 (A. Semtner, personal communication).

31.3.3 Sigma-Coordinate Models

Governing Equations 31.1 to 31.5 can be cast in a bottom-topography-following coordinate system by defining a new variable $\sigma = (z-\eta)/(H+\eta)$ and transforming the equations to the new coordinate system [Blumberg and Mellor 1987]; (see also Kantha and Piacsek [1993] for the general orthogonal curvilinear coordinate form):

$$\frac{\partial \eta}{\partial t} + \frac{\partial (U_k D)}{\partial x_k} + \frac{\partial \omega}{\partial \sigma} = 0 \tag{31.17}$$

$$\frac{\partial (U_j D)}{\partial t} + \frac{\partial}{\partial x_k}(U_k U_j D) + \frac{\partial}{\partial \sigma}(\omega U_j) + f\varepsilon_{j3k} U_k D = -D\frac{\partial P}{\partial x_j} + D\frac{\partial \Phi}{\partial x_j} + \frac{\partial}{\partial \sigma}\left(\frac{K_M}{D}\frac{\partial U_j}{\partial \sigma}\right) + DF_j \tag{31.18}$$

$$\frac{\partial P}{\partial \sigma} = -\frac{\rho}{\rho_0} g D \tag{31.19}$$

$$\frac{\partial (\Theta D)}{\partial t} + \frac{\partial}{\partial x_k}(U_k \Theta D) + \frac{\partial}{\partial \sigma}(\omega \Theta) = \frac{\partial}{\partial \sigma}\left(\frac{K_H}{D}\frac{\partial \Theta}{\partial \sigma}\right) + DS_\Theta + DF_\Theta \tag{31.20}$$

$$\frac{\partial (SD)}{\partial t} + \frac{\partial}{\partial x_k}(U_k S D) + \frac{\partial}{\partial \sigma}(\omega S) = \frac{\partial}{\partial \sigma}\left(\frac{K_H}{D}\frac{\partial S}{\partial \sigma}\right) + DF_S \tag{31.21}$$

where $D = H + \eta$, the total depth of the water column, and ω is the pseudo vertical velocity in the new coordinate system, zero at the ocean surface ($\sigma = 0$) and the bottom ($\sigma = -1$).

These equations, along with corresponding conservation relations for turbulence quantities, form the basis of the popular **sigma-coordinate** Princeton model developed by George Mellor's group at Princeton University [Blumberg and Mellor 1987; see also Mellor 1991, Kantha and Piacsek 1993]. In this coordinate system, the number of levels is the same everywhere in the ocean, irrespective of the depth of the water column. It is therefore possible to resolve the bottom boundary layer where needed. This set of equations is best suited to modeling the shallow coastal oceans, although there is no inherent barrier to its application to deep basins. The principal problem is in applying it over sharply changing topography such as the continental slope separating the shelf from the deep basin. Here, unless the topographic gradients are suitably reduced by a nonlinear smoother, the errors in the calculation of pressure gradients induced by horizontal gradients of density can lead to spurious along-slope currents [Haney 1991]. While the problem due to strong topographic changes manifests itself in one form or another in all ocean models, the problem is particularly serious in sigma-coordinate models.

Many applications of this model can be found in the literature (for example in the *Journal of Physical Oceanography* and the *Journal of Geophysical Research, Oceans*). An application of a modified version developed at the University of Colorado, incorporating an improved mixed-layer formulation [Kantha and

Clayson 1994] and involving assimilation of altimetric data, is given in Section 31.4. This version has also been converted to CM5 and applied to the Straits of Sicily, and its Cray T3-D version is being applied to the North Pacific Ocean.

31.3.4 Layered Models

In layered models, the ocean is divided into several (N) layers in the vertical, and Equation 31.1 to Equation 31.3 are integrated over each layer ($n = 1, \ldots, N$) to obtain expressions for the thickness of and velocity in each layer. For example, Wallcraft [1991] obtains

$$\begin{aligned}
&\frac{\partial h^n}{\partial t} + \frac{\partial}{\partial x_k}\left(h^n U_k^n\right) = w^n - w^{n-1} \\
&\frac{\partial\left(h^n U_j^n\right)}{\partial t} + \left[\frac{\partial}{\partial x_k}\left(h^n U_k^n\right) + U_k^n \frac{\partial}{\partial x_k}\right] U_j^n + f\varepsilon_{j3k} h^n U_k^n \\
&\quad = -h^n \sum_{k=1}^{N} G_k^n \frac{\partial}{\partial x_k}\left(h^n - h_0^n\right) + \left(\tau_j^{n-1} - \tau_j^n\right) + A_M \frac{\partial^2}{\partial x_k \partial x_k}\left(h^n U_j^n\right) \\
&\qquad + \max\left(0, -w^{n-1}\right) U_j^{n-1} + \max(0, w^k) U_j^{n+1} - \left[\max(0, -w^n)\right. \\
&\qquad + \left.\max(0, w^{n-1})\right] U_j^n + \max(0, -c_{de} w^{n-1})\left(U_j^{n-1} - U_j^n\right) \\
&\qquad + \max(0, -c_{de} w^n)\left(U_j^{n+1} - U_j^n\right)
\end{aligned} \tag{31.22}$$

where h^n is the thickness and U_j^n the velocity of the nth layer, w^k is the vertical velocity at the kth interface, and h_0 is the layer thickness at rest. The Nth layer contains the model basin topography, and its thickness is the total depth of the water column minus the sum of the thicknesses of the remaining layers. Finally,

$$\begin{aligned}
G_k^n &= \begin{cases} g, & k \geq n \\ g\left[1 - \dfrac{\rho^n - \rho^k}{\rho_0}\right], & k < n \end{cases} \\
\tau_j^n &= \begin{cases} \tau_w, & n = 0 \\ c_{dn}\left|U_j^n - U_j^{n+1}\right|\left(U_j^n - U_j^{n+1}\right), & n = 1, \ldots, N-1 \\ c_{db}\left|U_j^N\right| U_j^N, & n = N \end{cases}
\end{aligned} \tag{31.23}$$

The factor c_d is the drag coefficient, c_{de} is the drag due to entrainment of fluid from one layer to the adjacent one, τ_w is the wind stress, and ρ^n is the density of the nth layer. Note that the layer densities do not change with time, only their thickness does at each model grid point. The conditional statements have to do with entrainment and detrainment at each interface between two adjacent layers, the details of which can be found in Wallcraft [1991].

The thinning of a layer to vanishing thickness is a major problem in layered models that leads to numerical difficulties. The traditional solution has been to make each layer thick enough, but this distorts the representation of the oceanic vertical structure. An alternative solution is to entrain fluid into the thinning layer from below to thicken it. Such entrainment has to be balanced by global detrainment in the layer so as to keep the density of each layer constant in space and time. For details of this and the model numerics, see Wallcraft [1991].

It is essential to select the number and rest thicknesses of layers carefully in layered models. Since topographic variations are contained in the bottommost layer only, these models are generally incapable of simulating circulation in coastal and shallow seas. They are, however, excellent at capturing the important lowest-order dynamics of the basin circulation and are therefore widely used for process-oriented studies. They are also being increasingly used for a variety of applications. One example is the six-layer, $\frac{1}{8}^\circ$ global

model at the Naval Research Laboratory at Stennis Space Center, Mississippi, the SSH from which is shown in two parts, the Atlantic and Indian Oceans in Figure 31.3a and the Pacific Ocean in Figure 31.3b. Realistic depiction of mesoscale activity, especially in regions of strong ocean currents — such as the Gulf Stream in the Atlantic, Kuroshio in the Pacific, the Brazil/Malvinas Current off Brazil, the Agulhas Current off Africa, and the Circumpolar Current around the continent of Antarctica — are noteworthy. The SSH variability from a layered model like this, driven by synoptic winds from a NWP center such as Fleet Numerical Meteorology and Oceanography Center, compares well with the variability indicated by altimeters such as the U.S. Navy's GEOSAT.

A simple subset of the layered model is the so-called reduced-gravity model (also called $1\frac{1}{2}$-layer model), where the water column is assumed to consist of two layers: an active top layer of thickness H and a quiescent bottom layer of infinite thickness, with a density interface between the two of intensity $\Delta\rho$. It is remarkable that this very simple model often captures the essential dynamics of the circulation; for example, a reduced-gravity model of the Gulf of Mexico demonstrated conclusively that it is the instability of the Loop Current that is responsible for the shedding of the Loop Current eddies [Hurlburt and Thompson 1980]. The governing equations are identical to the barotropic Equation 31.15, except that the gravity parameter g is replaced by $g' = g\,(\Delta\rho/\rho_0)$, the reduced gravity (whose value is two orders of magnitude smaller than g; hence the name reduced-gravity model), with H now denoting the rest thickness of the upper layer and η denoting the deflection of the interface.

31.3.5 Isopycnal Models

Isopycnal models are similar to the layered models discussed above but are fully dynamical and thermodynamic. Despite the numerical problems associated with surfacing and vanishing of layers, they are well suited to simulate basin dynamics. Considerable progress has been made over the last decade in isopycnal modeling, and with the inclusion of adequate upper-mixed-layer physics they are also becoming quite practical. Examples of applications can be found in Oberhuber [1993] and Bleck and Smith [1990]. Since they principally deal with isopycnals (surfaces of equal density) and do not consider temperature and salinity separately, but instead treat density as the prognostic variable, they are not well suited to handling situations where temperature and salinity must be computed separately. A linear equation of state and identical diffusion characteristics for temperature and salinity are implicit in these models. This is valid over a majority of the global oceans, if one excludes regions such as those near river outflows and sea-ice formation.

31.3.6 Data Assimilation

Inevitable errors in initial conditions and imperfect parametrization of physical processes make a model ocean diverge rapidly from the real ocean. This is simply due to the extreme sensitivity of this system to even minute changes in initial conditions, typical of chaotic nonlinear systems. It is therefore essential to employ observational data in ocean models to retain the modeled ocean state close to the real state. The situation is no different from that in modeling the state of the atmosphere for NWP purposes, except that the time scales for loss of predictability is weeks for the oceans compared to days for the atmosphere. The process of employing observed data from the real ocean (atmosphere) to keep the modeled ocean (atmosphere) realistic is called data assimilation [for example, Anderson and Moore 1986] and consists of combining the modeled fields with observed data at various points in the domain to produce the best possible estimate of the real state of the ocean over the entire model domain. Exactly how this is best done has been the subject of considerable research in the atmospheric community [Bengtsson et al. 1981] over the past few decades, and more recently in the oceanic community as well [Haidvogel and Robinson 1989].

NWP centers use predominantly the so-called analysis–forecast cycle of assimilation. Here, the current state of the modeled atmosphere as predicted by the previous forecast is combined with observations of the atmosphere by radiosondes and surface stations all over the world, by an analysis–initialization process, to produce initial fields of various model variables suitable for describing the initial state for the next model

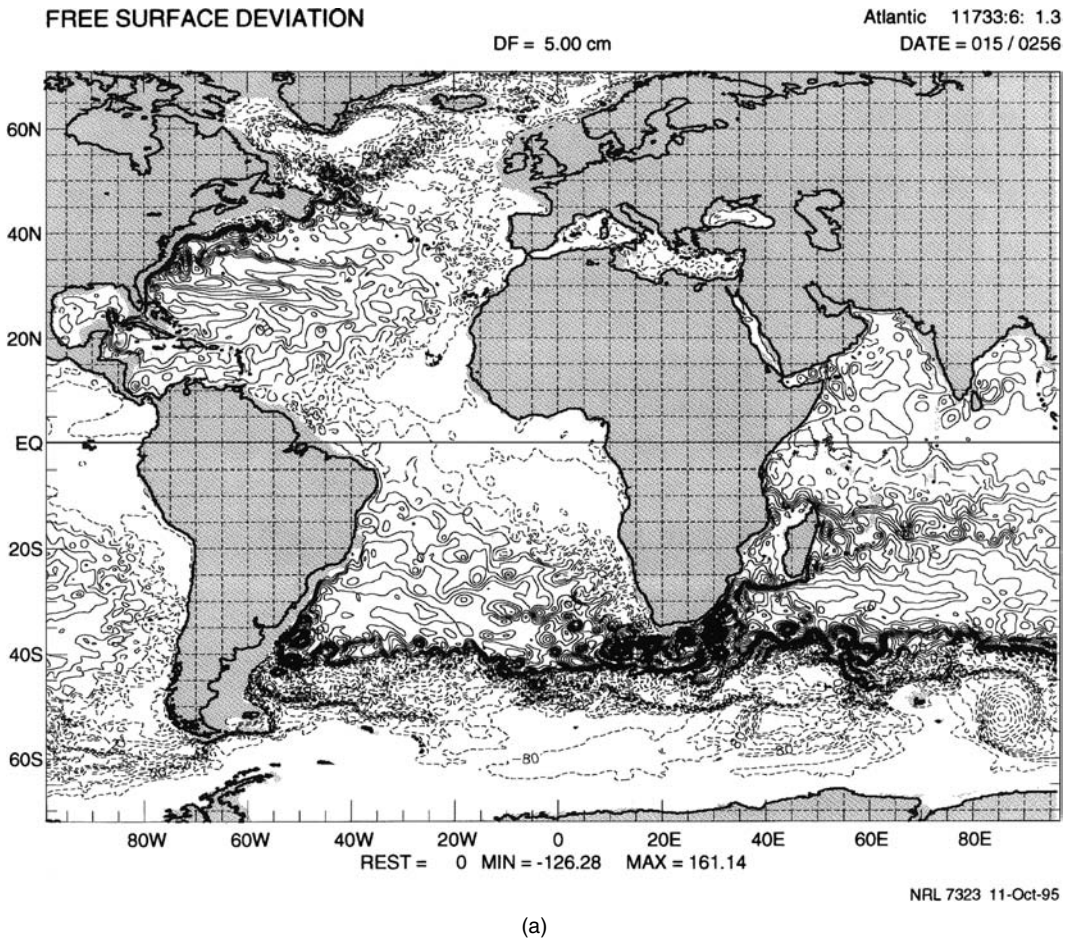
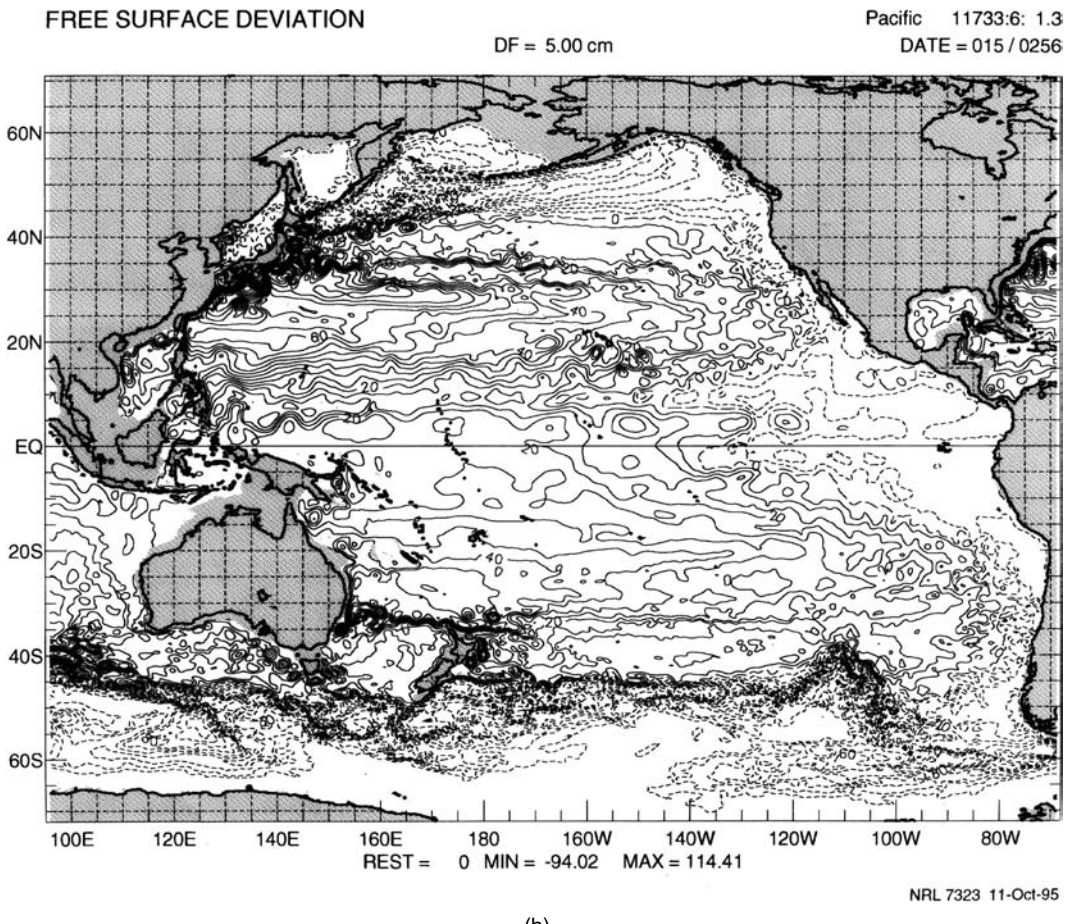


FIGURE 31.3 Sea surface from the six-layer Naval Research Laboratory $\frac{1}{8}^{\circ}$ global model for the Atlantic and the Indian Oceans (a) and for the Pacific Ocean (b). Note the eddy-resolving capability of this model displayed in the realistic mesoscale activity in regions of strong currents such as the Aghulas around Africa.

forecast. The forecast skill depends very much on the accuracy of the initial state so derived, since errors in the latter tend to get magnified with time during the forecast.

Another possible assimilation method is the so-called continuous assimilation [Bengtsson 1981], where a numerical model is kept running and current by assimilation of observed data as and when they become available. A forecast can then be initiated by a similar model running forward *free* (without any data assimilation). The principal advantage of this method over the analysis–forecast cycle is that the model derives benefit from all past data as opposed to a single set of observations at a particular time. Also, the shock of data insertion due to inevitable mismatch between the model and observed states is less severe. Since such a mismatch can lead to severe noise superimposed on the true state of the forecast atmosphere (ocean), often making the forecast worthless, considerable effort has been expended in devising means to minimize such a mismatch, resulting in a procedure called *initialization* in NWP terminology. Continuous assimilation tends to reduce this shock and is therefore often preferable. The reader is referred to Bengtsson et al. [1981] and Haidvogel and Robinson [1989] for a discussion of assimilation philosophies.

The method of combining data into a model can vary from the simplest one (called data insertion), in which the model-predicted values are just replaced by observed values, to Kalman filters [Gelb 1988] (which blend the model and observed values optimally, taking into account the model error and observational



(b)

FIGURE 31.3 *continued*

error statistics), adjoint techniques [Thacker and Long 1987], and variational methods [Derber and Rosati 1989]. It is also possible to use nudging techniques in which appropriate Newtonian damping terms that damp the variable to the observed value with a predetermined time scale are introduced into the governing equations. The most commonly employed method is optimal interpolation [see Choi and Kantha 1995, for example], since methods such as Kalman filters and adjoint techniques are computationally expensive and at present still impractical for applications in NWP and ocean prediction.

It is beyond the scope of this article to go into details of data assimilation methods. Instead, the reader is referred to the above references (and more recent work in the literature, especially on NWP), with a reminder that most assimilation methods replace the model-predicted values by a weighted combination of model-predicted value and observed values during the assimilation step, with the weight either determined a priori by statistical methods such as optimal interpolation or updated at each assimilation step by a method such as Kalman filtering. For examples of oceanic data assimilation, the reader is referred to Derber and Rosati [1989], Glenn and Robinson [1995], and Choi and Kantha [1995].

31.3.7 Computational Issues

Ocean models make a large demand on computer resources, CPU time, core memory, and disk storage, because ocean eddies are much smaller than weather systems, and the resolution needed is therefore much

finer. Fine resolution also forces one to take smaller time steps in explicit models on account of CFL constraints. Even in "implicit" ocean models, the advection terms are treated explicitly, thus imposing a time-step limitation.

For explicit free-surface models, the time step is limited by the step of the fast-moving surface gravity waves, and one has to take a large number of small time steps to integrate over a simulation or forecast period (mode splitting helps alleviate this problem). For implicit ocean models, which filter out these gravity modes, the CPU-time requirement is governed by the rate of convergence of the iterative method used to solve the resulting Poisson equation.

Explicit model codes are usually readily vectorizable and parallelizable and generally need few additional arrays to store the auxiliary variables that may be needed to speed up the computations. In contrast, the vectorization/parallelization of the Helmholtz solvers associated with implicit codes is usually a nontrivial problem and, for some schemes that have been used up to now on serial machines, not at all feasible. The extra work resulting from the iterative or matrix inversion solution can often increase the total CPU time so that it is comparable to, or even exceeds, that for explicit codes, especially on vector/parallel computers. In addition, there are almost always extra arrays needed during this stage of the computations. The two- or four-color versions of the successive overrelaxation (SOR) and the conjugate-gradient method are two techniques that are well suited to vectorization and parallelization in implicit codes.

In the early days of supercomputers, the core memory available was usually so small that all the arrays needed for computations in ocean models, especially global ones, would not fit within the core, and elaborate methods were employed to make efficient use of high-speed disks to transfer arrays into and out of core as needed. GFDL models (MOM2 for example) still retain such an architecture. With high-speed memory becoming much cheaper, modern supercomputers have core memories measured in gigawords (Gw), and many ocean models can now reside in memory, although the need for out-of-core models has not totally vanished, especially for very high resolutions and global coverage. In-core models such as the Los Alamos POPS are, however, better suited to efficient massive parallelization than the out-of-core ones such as MOM, because of the considerable disk I/O involved.

Disk/tape storage requirements for storing ocean model results are also often in tens to hundreds of gigabytes and depend on the length of the simulation and on how often and how many variables are required to be stored for later analyses. Disk storage and postprocessing requirements often constrain the temporal resolution and the details of the analyses carried out on the results of an ocean model.

Data-assimilative ocean models require even more resources than the free-running ones, with the additional memory and CPU-time requirements depending very much on the method of assimilation. It is not unusual for assimilation to more than double the CPU time requirements, even for simple OI-type schemes. Methods such as Kalman filters and adjoint methods are even more demanding. Generally, data assimilation on massively parallel computers requires considerable investment of time and effort for efficient implementation. We will give some typical CPU-time and memory requirements for large ocean models and for diverse computers, to cover a spectrum of configurations and to familiarize the reader with resource requirements of computational ocean modeling. The $\frac{1}{8}^\circ$ six-layer NRL global model ($2051 \times 1145 \times 6$ grid) requires 1.8 Gw of memory and 2 CPU hours per month of simulation on a 256-node CM5-E. The $\frac{1}{16}^\circ$ Pacific model ($1977 \times 1313 \times 6$ grid) requires 385 Megawords (Mw) and 17 single-processor CPU hours per month on a Cray C-90. The $\frac{1}{5}^\circ$ global explicit barotropic tidal model discussed in Section 31.3 under "Barotropic Models" (1801×729 grid) employs a time step of 13 s, assimilates 4000 data points every time step, and requires 65 Mw of memory and 22 single-processor CPU hours for a 10-day simulation on a 16-processor Cray C-90, assimilating 4000 data points every time step. A 15-level northern hemisphere Arctic ice–ocean model ($360 \times 360 \times 15$ grid) requires 40 Mw of memory and 25 CPU hours per month on a Cray C-90. A 30-level sigma-coordinate model of the eastern Pacific ($163 \times 229 \times 30$ grid) requires 42 Mw of memory and 4 CPU hours per month on Cray C-90. The small Gulf of Mexico sigma-coordinate nowcast–forecast model ($85 \times 86 \times 22$ grid) discussed in the next section requires 30 Mw of memory and 6 CPU hours for a month-long simulation in the nowcast mode, and 4 CPU hours in the forecast mode, the additional time requirements for the simple OI-based data assimilation being in this case about 50%. An idea of the storage requirements can be obtained from the fact that

even this small model required 2 Gbytes to store the model output at 5-day intervals for a 10-year-long simulation without any data assimilation, and postprocessing of this output required numerous hours on a powerful Sun Sparc workstation.

31.4 Nowcast/Forecast in the Gulf of Mexico (a Case Study)

An important application of ocean models is in prediction of the current (nowcast) and future (forecast) state of the ocean. Given the fact that more than 50% of the burgeoning human population lives within 100 miles of a coastline and hence uses/abuses the coastal oceans, such predictions, especially in the coastal and marginal seas, might be particularly useful for societal needs such as sea-level predictions, mapping of currents, and pollution tracking. We will provide one such example from a marginal semienclosed sea in the north Atlantic, the Gulf of Mexico. The offshore oil fields of this "mini-ocean basin" account for roughly half the U.S. domestic oil production, and the Louisiana–Texas (LATEX) continental shelf is dotted with thousands of oil platforms. Exploration and production are expanding steadily into deeper waters, waters as deep as 1000 m.

The major oceanic phenomenon in the Gulf is the so-called Loop Current variability. Every second, about 28 million cubic meters of subtropical waters enter the Gulf through the Yucatan Straits between Mexico and Cuba and leave it through the Florida Straits between Florida and Cuba to eventually become the Gulf Stream. The extent of penetration of this so-called Loop Current into the Gulf is highly variable. Occasionally the Loop Current becomes unstable and sheds off a huge anticyclonic (clockwise) eddy, anywhere from 100 to 350 km in diameter, that pinches off the Loop Current and moves into the western Gulf. This Loop Current eddy (LCE) is the principal mechanism for renewal of waters in the western Gulf [Hurlburt and Thompson 1980]. The path of LCEs is also highly variable, and occasionally a LCE traverses the Gulf in close proximity to the LATEX continental shelf. Because of the strong currents (often as much as 4 knots, 2 m s^{-1} in magnitude) associated with LCEs, this is the second major source of concern (the first being hurricanes in late summer and fall) to production and exploration activities in the Gulf. A capability to accurately forecast the movement of a LCE is valuable to the oil industry.

A forecast of the path an LCE takes is possible with the use of a numerical model of the Gulf. However, accurate information on the initial location of the LCE once it is shed and the corresponding Gulf-wide oceanic state is crucial to the forecast skill. An accurate nowcast is therefore essential, and this requires a data-assimilative numerical model. Since in situ data, even in the Gulf, are sparse and often nonexistent, remotely sensed data need to be relied upon for this purpose. Since the sea surface temperature from IR sensors is not always useful in locating a LCE (especially in summer) and since altimetry can almost always detect such an eddy if it happens to straddle its track, altimetric SSH anomalies can be assimilated into the ocean model to provide a reliable nowcast of not only the eddy location but also the initial state of the Gulf. Forecasts can then be made from this nowcast and the path of the LCE predicted.

The methodology employed by Choi and Kantha [1995] for producing a nowcast (in a hindcast mode, that is, prediction of past events for which data are available for verification) is called the *continuous* or *four-dimensional* assimilation method and has its origins in NWP. The model is run from a time in the past to the present, assimilating altimetric data track by track [see Choi and Kantha 1995 for details]. Altimetric SSH anomalies derived from NASA/CNES TOPEX and ESA ERS1 altimeters are converted to anomalies in the temperature in the water column and assimilated into the model using simple optimal interpolation. In the particular example shown here, the model was run starting at the beginning of January 1993 (day 1 corresponds to Jan. 1) to produce a nowcast for day 240, at which a LCE pinches off and separates from the Loop Current. The model is then run forward free (without assimilation) to produce a forecast over the next 40 days, assuming nonchanging winds over the period. The forecast skill is assessed by comparison with the nowcast, which was also carried out over the rest of 1993.

Figure 31.4 shows a comparison of the forecast and nowcast SSH fields over the Gulf at days 260 (top) and 280 (bottom). The forecasts are shown on the left and the corresponding nowcasts on the right.

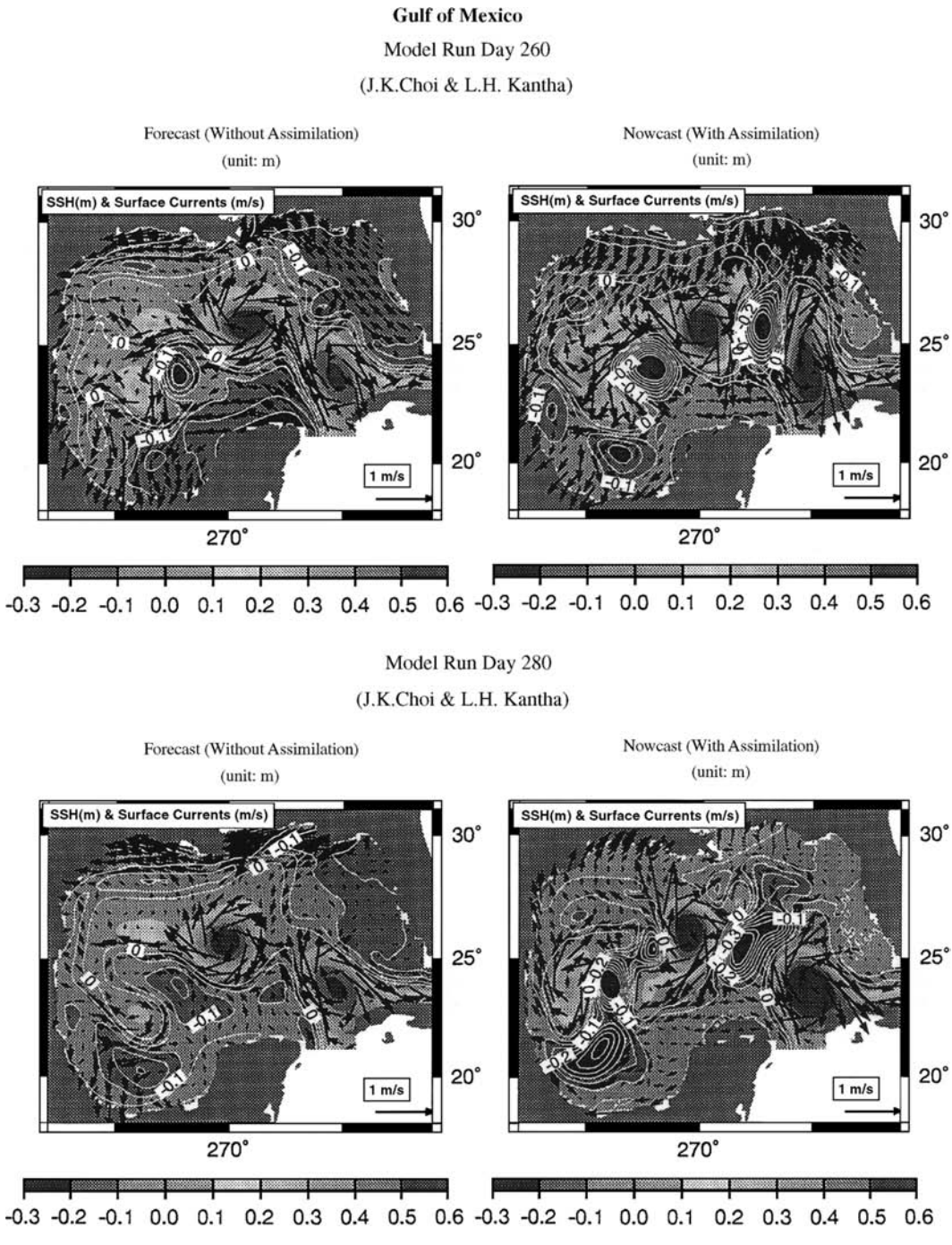


FIGURE 31.4 The sea surface elevation and currents from the forecast (left) and the nowcast (right) at days 260 (top) and 280 (bottom) from a three-dimensional circulation model of the Gulf of Mexico assimilating altimetric data from TOPEX and ERS1. The forecast was started at day 240. Compare the forecast with the corresponding nowcast to assess the model skill.

There is a close correspondence between forecast LCE position and the nowcast position, suggesting that the LCE path is being predicted reasonably well. The error, however, between forecast and nowcast LCE positions is larger at day 280 than at day 260. This particular experiment suggests that the forecast has some skill to about 30 days or so, beyond which the predicted path (forecast) deviates increasingly from the actual path (nowcast for the corresponding day). Since altimetric data are available within several days of their collection by the sensor, this experiment suggests that with some skill forecasts can be made two to three weeks in advance. If this is proven correct, this nowcast–forecast capability might be useful to drilling/exploration activities in the Gulf. It is in applications such as this that an ocean model, acting in concert with routine ocean monitoring via satellite-borne sensors, can prove useful.

31.5 Research Issues and Summary

We have provided a thumbnail sketch of ocean modeling as it is practiced today. As we said earlier, the field has undergone a phenomenal growth in recent years, and it is impossible to do justice to the subject in a short review like this. The reader is encouraged to pursue a particular model or approach of interest via the references cited.

The major issue in ocean modeling is the dearth of data for model initialization, forcing, assimilation, and of course verification or skill assessment. In situ data are rather sparse and, given the cost of ship time, likely to remain so. Therefore, increasing reliance will be placed on remote sensing to fill in gaps. However, this approach itself has limitations, and it is not clear what might fill the gap. Smart autonomous vehicles, a product of the Cold War, roaming the world oceans, and buoys sprinkled into the global oceans, telemetering data via communication satellites, may one day provide more in situ data than we currently acquire. Combined with multiteraflop computing capabilities of the coming century, an ocean observing and monitoring system consisting of satellites, moored arrays, buoys, and autonomous vehicles might one day finally enable us to set up realistic ocean prediction systems to satisy the needs of the coming generations. Ocean modeling will play a central role in all this.

Acknowledgments

Lakshmi Kantha acknowledges with pleasure the support provided by The Minerals Management Service of the Department of the Interior through an interagency agreement with the U.S. Navy through contract N00014-92C-6011, administered by Walter Johnson of MMS and Donald Johnson of the Naval Research Laboratory. Lakshmi Kantha was also supported by the NOMP program of the Office of Naval Research under contract N00014-95-1-0343, administered by Tom Curtin, and by the Coastal Sciences Section of the Office of Naval Research under contract N00014-92-J-1766, administered by Thomas Kinder.

Defining Terms

Altimeter: A microwave device measuring the time delay between an emitted microwave signal and its return by reflection from the sea surface. When the position of the instrument in space is independently determined, it enables sea surface topography to be measured to an accuracy of a few centimeters along the satellite track.

Baroclinic: Conditions in which the vertical shear is generated because of the horizontal gradients of density.

Barotropic: Conditions in which there are no variations in currents in the vertical direction.

Coamplitude and cophase: Lines of maximum tidal amplitude and lines of the time of occurrence of maximum tide, referred to either local or universal time.

Coriolis force: A fictitious force needed to allow for the noninertial nature of a rotating coordinate system.

Data assimilation: The process of blending observational data into numerical models.

El Niño: A frequent phenomenon in the tropical Pacific, occurring at 3–7-year intervals, when the eastern Pacific gets anomalously warm and sets off changes in the tropical atmosphere that affect weather all over the globe.

Gravimetry: The science of precise measurement of the earth's gravity.

Hindcast: A forecast excercise conducted for a period in the past to take account of the availability of accurate observational data for forcing, assimilation, and verification.

Inverse barometer effect: The effect where the changes in the atmospheric pressure are compensated exactly by the ocean by inverse changes in its height so that no oceanic motions are induced.

Isopycnal: A surface on which the density is constant.

Kelvin waves: Waves that run along the ocean margins (with the coast to the right in the northern hemisphere) at the speed of the shallow-water gravity wave. These waves are important for oceanic adjustment to changing surface forcing.

Nowcast: An estimate of the present state, often by an optimal blend of model and data.

Potential temperature: The temperature attained by a parcel of water brought adiabatically to a reference depth.

Reynolds averaging: The process of obtaining equations for mean quantities in a turbulent flow by considering each quantity to consist of a mean and a fluctuating component and taking averages over time or realizations.

Sigma coordinates: A coordinate system where the vertical coordinate is normalized by local depth; it is bottom-fitting or topographically conformal.

Synoptic forcing: Multihourly forcing from atmospheric models run at NWP centers, obtained in the past from a synopsis of weather charts.

Western intensification (boundary current): Strong currents found at the western boundaries of the ocean basins or eastern sides of continents because the effect of Earth's rotation variation with latitude (the so-called β-effect).

References

Andersen, O. B., Woodworth, P. L., and Flather, R. A. 1995. Intercomparison of recent ocean tidal models. *J. Geophys. Res.* 100:25261–25282.

Anderson, D. L. T. and Moore, A. M. 1986. Data assimilation. In *Advanced Physical Oceanographic Numerical Modelling*, J. J. O'Brien, Ed., pp. 437–464. Reidel, Dordrecht.

Bengtsson, L., Ghil, M., and Kallen, E., ed. 1981. *Dynamic Meteorology: Data Assimilation Methods*, p. 330. Springer–Verlag, New York.

Bleck, R. and Smith, L. T. 1990. A wind-driven isopycnic coordinate model of the north and equatorial Atlantic Ocean. Part I: Model development and supporting experiments. *J. Geophys. Res.* 95:3273–3285.

Blumberg, A. F. and Kantha, L. H. 1985. Open boundary conditions for circulation models. *J. Hydraulic Eng.* 111:237–255.

Blumberg, A. F. and Mellor, G. L. 1987. A description of a three-dimensional coastal ocean circulation model. In *Three-dimensional Coastal Ocean Models*, N. Heaps, Ed., pp. 1–16. American Geophysical Union, Washington, DC.

Bryan, K. 1969. A numerical model for the study of the circulation of the world oceans. *J. Comput. Phys.* 4:347–359.

Choi, J.-K. and Kantha, L. H. 1995. A nowcast/forecast experiment using TOPEX/ Poseidon and ERS-1 altimetric data assimilation into a three-dimensional circulation model of the Gulf of Mexico. Abstract, XXI IAPSO Meeting, Hawaii, Aug. 5–12.

Cox, M. D. 1985. An eddy-resolving numerical model of the ventilated thermocline. *J. Phys. Oceanogr.* 15:1312–1324.

Derber, J. and Rosati, A. 1989. A global oceanic data assimilation system. *J. Phys. Oceanogr.* 19:1333–1347.

Desai, S. D. and Wahr, J. M. 1995. Empirical ocean tide models estimated from TOPEX/POSEIDON altimetry. *J. Geophys. Res.* 100:25205–25228.

Dietrich, D. E., Marietta, M. G., and Roach, P. J. 1987. An ocean modeling system with turbulent boundary layers and topography. *Int. J. Numer. Methods Fluids* 7:833–855.

Dukowicz, J. K. and Smith, R. D. 1994. Implicit free-surface model for the Bryan–Cox–Semtner ocean model. *J. Geophys. Res.* 99:7991.

Dukowicz, J. K., Smith, R. D., and Malone, R. C. 1993. A reformulation and implementation of the Bryan–Cox–Semtner ocean model on the Connection Machine. *J. Atmos. Ocean. Technol.* 10:195.

Gelb, A., ed. 1988. *Applied Optimal Estimation*, p. 374. MIT Press, Cambridge, MA.

Gill, A. E. 1982. *Atmosphere–Ocean Dynamics*. p. 666. Academic Press, New York.

Glenn, S. M. and Robinson, A. R. 1995. Verification of an operational Gulf Stream forecasting model. In *Quantitative Skill Assessment for Coastal Ocean Models*, D. R. Lynch and A. M. Davies, Eds., pp. 469–499. American Geophysical Union, Washington, DC.

Haidvogel, D. B. and Robinson, A. R. 1989. In *Data Assimilation*, Special issue, *Dyn. Atmos. Oceans*. 13:171–515.

Haney, R. L. 1991. On the pressure gradient force over steep topography in sigma-coordinate ocean models. *J. Phys. Oceanogr*. 21:610–619.

Heaps, N., ed. 1987. *Three-Dimensional Coastal Ocean Models*. p. 208. American Geophysical Union, Washington, DC.

Hellerman, S. and Rosenstein, M. 1983. Normal monthly wind stress over the world ocean with error estimates. *J. Phys. Oceanogr*. 13:1093–1104.

Hibler, W. D., III and Bryan, K. 1987. Diagnostic ice–ocean model. *J. Phys. Oceanogr*. 17:987–1015.

Holland, W. R. 1986. Quasi-geostrophic modeling of eddy-resolved ocean circulation. In *Advanced Physical Oceanographic Numerical Modeling*, J. J. O'Brien, Ed., pp. 203–231. Reidel, Dordrecht.

Hurlburt, H. E. and Thompson, J. D. 1980. A numerical study of Loop Current intrusions and eddy-shedding. *J. Phys. Oceanogr*. 10:1611.

Kantha, L. H. 1995. Barotropic tides in the global oceans from a nonlinear tidal model assimilating altimetric tides. 1. Model description and results. *J. Geophys. Res*. 100:25283–25308.

Kantha, L. H., Blumberg, A. F., and Mellor, G. L. 1990. Computing phase speeds at an open boundary. *J. Hydraulic Eng.* 116:592–597.

Kantha, L. H. and Clayson, C. A. 1994. An improved mixed layer model for geophysical applications. *J. Geophys. Res*. 99:25235–25266.

Kantha, L. H. and Piacsek, S. A. 1993. Ocean models. In *Computational Science Education Project,* Oak Ridge Nat. Lab., Dept. Energy Rep. CSEP.

Kantha, L., Whitmer, K., and Born, G. 1994. The inverted barometer effect in altimetry: a study in the North Pacific. *TOPEX/Poseidon Res. News* 2:18–23.

Killworth, P. D., Stainforth, D., Webb, D. J., and Paterson, S. M. 1991. The development of a free surface Bryan–Cox–Semtner ocean model. *J. Phys. Oceanogr*. 21:1333–1348.

Kowalik, Z. and Murty, T. S. 1993. *Numerical Modeling of Ocean Dynamics*, p. 481. World Scientific, Singapore.

Le Provost, C., Genco, M. L., Lyard, F., Vincent, P., and Canceil, P. 1994. Spectroscopy of the world tides from a finite element hydrodynamical model. *J. Geophys. Res*. 99:24777–24797.

Levitus, S. 1982. Climatological atlas of the world ocean. *NOAA Professional Paper* 13, Geophys. Fluid Dyn. Lab., Princeton, NJ. 173 pp.

Madala, R. V. and Piacsek, S. A. 1977. A model for baroclinic oceans. *J. Comput. Phys*. 22:167.

Mellor, G. L. 1991. User's guide for a three-dimensional, primitive equation, numerical ocean model. *Princeton University Rep.* 91. 35 pp.

Mellor, G. L. and Yamada, T. 1982. Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys*. 20:851–875.

Mesinger, F. and Arakawa, A. 1976. *Numerical Methods Used in Atmospheric Models*, Vol. 1. Global Atmospheric Research Program Publication 17. 64 pp.

Metzger, E. J., Hurlburt, H. E., Kindle, J. C., Serkes, Z., and Pringle, J. M. 1992. Hindcasting of wind-driven anomalies using a reduced-gravity global ocean model. *Mar. Technol. Soc. J.* 26:23–32.
Oberhuber, J. M. 1993. Simulation of the Atlantic circulation with a coupled sea ice–mixed layer–isopycnal general circulation model. Part I: Model description. *J. Phys. Oceanogr.* 23:808–829.
O'Brien, J. J. 1985. *Advanced Physical Oceanographic Numerical Modeling*. Reidel, New York.
Orlonski, I. 1976. A simple boundary condition for unbounded hyperbolic flows. *J. Comput. Phys.* 21: 251–269.
Pond, S. and Pickard, G. L. 1979. *Introductory Dynamical Oceanography*, 2nd ed. p. 329. Pergamon Press, New York.
Ponte, R. M. 1994. Understanding the relation between wind driven sea level variability and atmospheric pressure. *J. Geophys. Res.* 99:8033–8040.
Roed, L. P. and Cooper, C. K. 1986. Open boundary conditions in numerical ocean models. In *Advanced Physical Oceanographic Numerical Modeling*, J. J. O'Brien, Ed., pp. 411–436. Reidel, Dordrecht.
Schwiderski, E. W. 1980. On charting global ocean tides. *Rev. Geophys.* 18:243–268.
Semtner, A. J. 1986. Finite-difference formulation of a world ocean model. In *Advanced Physical Oceanographic Numerical Modeling*, J. J. O'Brien, Ed., pp. 187–202. Reidel, Dordrecht.
Semtner, A. J. 1995. Modeling ocean circulation. *Science* 269:1379–1385.
Semtner, A. J., Jr. and Chervin, R. M. 1992. Ocean general circulation from a global eddy-resolving model. *J. Geophys. Res.* 97:5493–5550.
Smagorinskiy, J. 1963. General circulation experiments with primitive equations: I. The basic experiment. *Mon. Weather Rev.* 91:99–164.
Smith, R. D., Dukowicz, J. K., and Malone, R. C. 1992. Parallel ocean general circulation modeling. *Phys. D* 60:38.
Thacker, W. C. and Long, R. B. 1987. Fitting dynamics to data. *J. Geophys. Res.* 93:1227–1240.
Wallcraft, A. J. 1991. The Navy layered ocean model users guide. *NOARL Rep.* 35, 21 pp.
Warren, B. A. and Wunsch, C., Eds. 1981. *Evolution of Physical Oceanography*. p. 623. MIT Press, Cambridge, MA.

Further Information

This review chapter has been necessarily sketchy. The reader is therefore encouraged to consult the various references cited for more details. The monograph on numerical ocean modeling edited by James O'Brien [1985] is still the best starting point, especially since the models described there have remained essentially unchanged, undergoing only small evolutionary changes such as adaptation to massively parallel computers and inclusion of better mixing algorithms. A good starting point in coastal ocean modeling is the American Geophysical Union (AGU) volume edited by N. Heaps [1987]. Kowalik and Murty [1993] is an excellent "cookbook" for details of numerics such as finite-differencing and the split-mode technique. Reference can be made to Haidvogel and Robinson [1989] for a good description of data assimilation methods. Textbooks by Pond and Pickard [1979] and Gill [1982] are good starting points for exploring the dynamics of the oceans. The Henry Stommel 60th Birthday volume on physical oceanography [Warren and Wunsch 1981] is a good followup.

There is no specific journal for ocean modeling; instead, modeling advances are published in journals such as the *Journal of Physical Oceanography* of the American Meteorological Society (AMS) and *Journal of Geophysical Research (Oceans)* of the American Geophysical Union (AGU). The *Journal of Hydraulic Engineering* of the American Society of Civil Engineers publishes modeling papers mostly related to coastal and estuarine studies. The *Journal of Continental Shelf Research* specializes in coastal research, including coastal modeling. Purely computational advances often appear in journals such as the *Journal of Computational Physics*. Semiyearly meetings of the AGU, meetings of the AMS and biennial Ocean Sciences, and quadrennial meetings of the International Union of Geodesy and Geophysics (IUGG) are examples of venues where latest advances in ocean modeling are presented and critiqued.

The GFDL *z*-level deep-water-basin model (ftp.gfdl.gov; directory pub/ GFDL_MOM3), Princeton sigma-coordinate shallow-water coastal model (ftp.gfdl.gov; directory pub/slm), and University of Miami isopycnal model (http://oceanmodeling.rsmas.miami.edu/micom) are all available. Readers are encouraged to offload the model codes and experiment with them. A good starting point for hands-on ocean modeling is the Ocean Models chapter of the Computational Science Education Project [Kantha and Piacsek 1993], available on the World Wide Web at http://csepl.phy.oral.gov/csep.html. It contains model code, graphics, animation, and exercises on simple ocean models that serve as a good introduction to the field.

32

Computational Chemistry

Frederick J. Heldrich
College of Charleston

Clyde R. Metz
College of Charleston

Henry Donato
College of Charleston

Kristin D. Krantzman
College of Charleston

Sandra Harper
College of Charleston

Jason S. Overby
College of Charleston

Gamil A. Guirgis
College of Charleston

32.1 Introduction

The use of computational methods in the study of chemistry touches upon every area of chemical inquiry. Indeed, the art and the science of computation are a natural fit with the study of chemistry. From the earliest times, beginning even with alchemy, chemists have used models to render comprehensible the abstract theories and concepts of their field. It is only logical, therefore, that chemists would use the power of modern computational methods to extend and explore their understanding of chemical compounds and processes.

Computational applications in chemistry are so vast and varied that it would be impossible to cover the entire field within the confines of this chapter. Instead, we will provide an overview of the types of computational applications in the field of chemistry, followed by a more detailed presentation in a few areas to show how chemists integrate computation into their discipline.

Anyone who has taken an introductory course in chemistry will remember using calculations to solve chemical problems — at least equilibrium, kinetics, and stoichiometry problems. Chemists have become adept at using computational tools to solve such mathematical problems and in using those tools to model,

and thereby test, their understanding of chemical phenomena and processes. Tools such as spreadsheets, math programs, graphing calculators, and iconic modeling programs have replaced the slide rule of 40 years ago. These tools bring greater predictive power, better understanding, and the potential to solve ever more complex problems.

Chemists describe compounds at the most fundamental level as a reflection of the nature of the atoms, the bonds between those atoms, and the electrons that comprise them. In fact, since Schrödinger's development of functions in the 1920s to describe electrons as waves, chemists have attempted to describe chemical compounds and processes, with increasing sophistication, in purely mathematical terms. The limitations in this approach are both theoretical (the conceptual framework for our understanding of chemistry is not perfect) and practical (the mathematical and computational tools are not perfect, either). Since this approach was initiated, however, great strides have been and continue to be made in both the theoretical and practical arenas.

This overview of computational chemistry begins with a presentation of how students are introduced to computation, followed by a description of how mathematical modeling is used to comprehend chemical processes that do not require detailed understanding of the chemicals involved. The chapter concludes with a description of how chemists use computational methods to understand the structure of compounds and the nuances of chemical reactions.

32.1.1 Underlying Principles

For many areas of computational chemistry, mathematicians, physicists, and chemists (such as Schrödinger, Hartree, and Pauling) laid the theoretical foundation in the 1920s and 1930s. The power of today's desktop computers allow experimental chemists to bring these principles into practice. As these theories are tested and as the comparison of experimental and computational results reveals the theories' limitations, advances in theory continue to be made.

32.2 Computational Chemistry in Education

In the chemistry classroom, computation ranges from various forms of modeling (molecular and mathematical modeling, solving complex simulation problems, and molecular animation) to text and class supplements (homework and testing, computational tools, demonstrations and animations, and interactive figures). Computation appears in the chemistry laboratory as prelaboratory assignments (discussion of theoretical concepts and the proper use of equipment), simulated experiments and instruments, and the use of computational tools for data analysis.

Through the *Journal of Chemical Education* (*JCE*), the Division of Chemical Education of the American Chemical Society publishes articles on the theory and application of computational chemistry, summarizes symposia from national meetings, and reviews software programs. In addition, the division makes available inexpensive, high-quality software to instructors and students from pre–high school through graduate school through *JCE Software*, Project SERAPHIM, and the various Web-based services available to *JCE* subscribers. Software capabilities have paralleled advances in operating systems, from various flavors of Apple II and MS/PC-DOS to Macintosh and Windows.

32.2.1 Journal of Chemical Education

For more than 75 years, *JCE* has served as the primary source of information for chemical educators. Today, *JCE* offers monthly published issues, software through *JCE Software*, online access, and various printed materials. In addition to the general articles in *JCE*, several feature columns are useful to computational chemistry:

Reviews of books, media, and software
Computer Bulletin Board
JCE Online

JCE Software
Molecular Modeling
Only@JCE Online, featuring *JCE* WebWare, Mathcad in the Chemistry Curriculum, and WWW site reviews
Teaching with Technology

The journal is fully searchable, with approximately 15 index keywords related to computational chemistry. To illustrate the current quality of coverage, a search using the keywords *computation chemistry* for the year 2002 resulted in nearly 200 hits.

Many general articles and Only@JCE Online features carry a symbol resembling the capital letter *W*. This symbol indicates that supplementary material (such as software, live spreadsheets and worksheets, additional data and exercises, laboratory instructions, animations, and video) is available to subscribers online. For example, one can link to the programs needed to analyze a kinetics simulation model for a drug poisoning victim, the software for finding the irreducible representations in a reducible representation for hybridization of orbitals and molecular vibrational motion, or Mathcad worksheets for Hückel theory calculations.

A relatively new addition to *JCE* is the regular feature JCE WebWare, which presents various Web-based applications suitable for computational chemistry in the laboratory, the classroom, or at home. Typically, the WebWare consists of small software programs, add-ins for spreadsheets and other standard programs, animations, movies, Java applets, or static and dynamic HTML pages. Recently, JCE WebWare included offerings for acid–base equilibria, nomenclature, games, chemical formatting add-ins for MS Word and Excel, spreadsheet analysis for first-order kinetics, a determinant solver for Hückel theory calculations, mechanism-based kinetics simulations, data analysis tools, and point group calculations. Each month, links are provided for fully interactive Chime-based models of some of the molecules discussed in the general articles in *JCE*.

Mathcad, a symbolic math software program, has become important in chemistry and, in particular, in physical chemistry. The *JCE* regular feature, Mathcad in the Chemistry Curriculum, presents abstracts of submitted documents that are useful for various chemistry courses. Recent worksheets include Hückel theory calculations, Bohr correspondence principle, NMR, modeling pH in natural waters, and variational treatment of a harmonic oscillator.

32.2.2 Project SERAPHIM

Project SERAPHIM began over 20 years ago as a clearinghouse for software and materials for computational chemistry. Membership fees support the clearinghouse, teacher workshops, and summer fellowships for software development. Software was distributed at a modest cost to members. A catalog listing several hundred different computer programs is available at the SERAPHIM Web site (http://ice.chem.wisc.edu/seraphim). These programs include

Databases — NMR library and experiments published in *JCE*
Data analysis tools — Significant figures, least squares analysis, dimensional analysis, spreadsheet calculations, and integration techniques
Chemical clip art
Teaching support — Quiz preparation and computer testing
Simulation problems — Wastewater treatment and water pollution
Tutorials and games
Laboratory — Experiment simulations, analytical techniques, and instrument interfacing

Core curriculum topics include states of matter (gases and crystals), nomenclature and formulas, stoichiometry (chemical equations, titrations, limiting reactants, and equilibria), thermochemistry, atomic structure (**electron configurations**, orbitals, and **SCF** calculations), chemical bonding (**molecular orbitals**, **HF** calculations, **Hückel calculations**, and **group theory**), spectroscopy (atomic, **NMR**, and **ESR**), dynamics (distribution of gas molecular speeds and chemical reaction kinetics), and descriptive chemistry

(acid–base, redox and electrochemistry, complexes, qualitative analysis, reaction prediction, organic molecules, biochemistry, polymers, and industrial chemistry). Currently, all Project SERAPHIM materials are available as free downloads from the Web site.

32.2.3 JCE Software

JCE Software resulted from collaboration between *JCE* and Project SERAPHIM with initial support from the Dreyfus Foundation. The motto of the journal is "*JCE Software* is not *about* software, it *is* software." Originally, the journal contained three series: for Apple II, Macintosh, and MS/PC-DOS users; later, a fourth series was added for Windows users. The software and corresponding printed materials were sent to subscribers twice a year. Currently, in addition to various video materials, *JCE Software* offers over 15 special issue software collections, covering laboratory and supplementary classroom materials for Macintosh and Windows users:

General and advanced chemistry collections — Student-designed collections featuring animations, simulations, and computational tools for acid–base chemistry, equilibria, spectroscopy, crystal structure, and **quantum mechanics**

Chemistry Comes Alive! collections — Movies, pictures, and animations of reaction types, stoichiometry, states of matter, thermodynamics and electrochemistry, organic chemistry and biochemistry, and laboratory techniques

Collections on specific topics — Periodic table, laboratory techniques, NMR, solid-state surfaces, material science, crystallography, and problem-based learning

Many of the software programs are Web-ready, and appropriate licensing is available for local intranets. Much of the older software for MS/PC-DOS is available to subscribers as free downloads at the *JCE* Web site.

32.3 Computational Aspects of Chemical Kinetics

Chemical reaction sequences (CRSes) are ubiquitous in chemistry and biochemistry. CRSes are used to describe models of phenomena as diverse as the sequence of elementary steps occurring in a single chemical or biochemical reaction, a metabolic sequence of reactions, and the complex chemical processes occurring in environmental systems, such as the atmosphere. It is almost always of interest to understand the evolution of these systems in time. Writing the differential rate equations for each elementary reaction in the sequence conveniently expresses the theoretical temporal evolution. If the rate constant for each elementary step is known, then, in principle, one has a complete description of the temporal evolution of that CRS. However, since experimentally one can usually measure the concentration of one or more of the species involved in a CRS at different times, comparing the theoretical model and the experimental data involves one of the following:

Differentiating the experimental data — Analysis of enzyme kinetics using the **Michaelis–Menton** equation

Integrating the coupled set of differential rate equations associated with the CRS — Determining the order of a chemical reaction by plotting ln([Reactant]) vs. time, 1/[Reactant] vs. time, etc.

A great deal of effort has gone into developing computational procedures that can convert the theoretical description of CRS into concentration vs. time information, which may then be compared directly to experimental results. There are two major approaches used to accomplish that goal: the numerical solution of differential equations and the **Monte Carlo** approach. Each has its own set of advantages and disadvantages, and software packages are available for each.

32.3.1 Numerical Solution of Differential Equations

CRS can be described by coupled sets of differential equations. Integrating these equations analytically is often not possible, so numerical techniques must be used. The various numerical techniques have been

described and sample programs have been presented in the literature (e.g., [Press et al., 1992]). Many CRSes of interest are **stiff**, that is, they contain rate constants that span many orders of magnitude. In order to analyze these systems effectively, implicit numerical methods, such as the one developed by Gear [Gear, 1971], must be used. The following programs, some of which may be downloaded free or for a modest fee, accomplish integration of coupled sets of stiff differential rate equations:

Gespasi [Mendes, 1993, 1997]
KINSIM [Frieden, 1993]
BerkeleyMadonna
Kintecus

Many of these software packages offer the capability to optimize the CRS under consideration. That is, the set of kinetic constants that bring the model in closest agreement with the kinetic data can be found [Mendes and Kell, 1998]. It is also possible to simulate stiff problems using Mathcad in conjunction with VisSim. Other software packages, which do not handle stiff differential equations (e.g., STELLA), have been used for less demanding applications. One of the most dramatic uses of stiff differential equation solvers to study CRS is the study of the ozone chemistry in the atmosphere. There, laboratory studies of individual atmospheric elementary reactions, coupled with atmospheric models, led to the decision to stop using CFCs. The 1995 Nobel lectures of Rowland, Molina, and Crutzen summarize this research [Crutzen, 1996; Molina, 1996; Rowland, 1996].

32.3.2 Monte Carlo Methods

Using an entirely different approach, researchers have developed methods to model the stochastic events of CRS rather than find numerical solutions to the coupled differential equations describing the CRS. An early report by Kibby [Kibby, 1969], followed by a complete theoretical exposition by Gillespie [Gillespie, 1976, 1977], describes **Monte Carlo** methods for simulating the time evolution of molecular events occurring in CRS. The probability of a particular reaction event occurring is the product of the intrinsic probability of such a reaction event (given by the rate constant) and the number of possible reaction events (given by the numbers of reacting molecules). The time interval is chosen in which the next reaction event is likely to occur, and then the particular reaction event that occurs is chosen from all possible reaction events. After adjusting the number of molecules for that event, the process is repeated.

This is how the actual stochastic events in a CRS are simulated. While the simulation can involve a very large number of events, stiff CRSes are simulated in exactly the same way as nonstiff CRSes. Furthermore, the method applies to very small-volume systems, such as a living cell or cellular organelle. This makes the method attractive to biochemists, in whose work it may not be appropriate to treat concentrations of molecules as a continuously varying quantity that changes deterministically over time [McAdams and Arkin, 1997]. A software package developed at IBM is available for free download that implements stochastic simulations of CRS.

32.4 Molecular Dynamics Simulations

In general chemistry, students are introduced to Dalton's intuitive picture of chemical reactions, in which atoms collide and rearrange to form new combinations. The computational method of molecular dynamics (MD) simulations maintains this classical view by treating atoms as billiard balls, with forces between them akin to springs [Garrison et al., 2000]. MD has a breadth of applications in chemistry, ranging from the determination of the lowest energy configuration of a protein to the study of mechanisms of chemical reactions. A distinct advantage of MD over methods based on electronic structure theory is that MD can be applied to large systems composed of thousands of atoms. Quantities can be calculated from the simulations, which can be compared with experimental values. Animations of atomic motions as a function of time are created, which can contribute mechanistic insights on a microscopic level unobtainable by experiment.

This section focuses on the application of MD simulations to study the high-energy bombardment of organic targets with atomic and polyatomic projectiles [Garrison et al., 2000; Zaric et al., 1998; Townes et al., 1999; Nguyen et. al., 2000], which is important in secondary ion mass spectrometry (**SIMS**). In SIMS, a primary ion beam bombards the surface with a low enough dose that each impact samples a fresh, undamaged portion of the surface. Secondary ions ejected from the surface are detected by a mass spectrometer.

SIMS is a widely used analytical technique, which is being developed for applications to molecular-specific imaging on a submicron scale [Berry et al., 2001]. Experiments have demonstrated that the secondary ion yield depends nonlinearly on the number of atoms in the projectile, and the nonlinear enhancement increases with the number of atoms [Van Stipdonk, 2001]. Therefore, polyatomic projectiles have the potential to improve significantly the sensitivity of SIMS. Sensitivity is the limiting factor in imaging applications, in which the maximum amount of analytical information must be obtained from a limited number of target molecules on the surface.

The objective of the simulations is to understand the mechanisms by which polyatomic projectiles enhance the secondary ion yield and to determine the optimum conditions for the use of polyatomic projectiles. The model systems used in these studies are composed of a thin organic film that is physisorbed to an atomic substrate. The simulations have compared the effect of Cu_n clusters with the same kinetic energy per atom [Zaric et al., 1998]. Simulations with SF_5 and Xe, which have the same mass, have been compared at the same bombarding energy. An illustration of the bombardment process is shown in Figure 32.1. Here, an SF_5 projectile impacts a monolayer of biphenyl molecules on a silicon substrate [Townes et al., 1999]. The energetic particle strikes the surface and dissipates its kinetic energy through the solid. Collision cascades develop, and molecules are lifted from the surface into the vacuum by the underlying substrate atoms.

32.4.1 The Methodology of Molecular Dynamics Simulations

A microcrystallite composed of N atoms is constructed that models the experimental system of interest. The nuclear motions of the atoms are assumed to obey the laws of classical mechanics:

$$\mathbf{F}_i = m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} \text{ which can be expressed as } \frac{d\mathbf{v}_i}{dt} = \frac{\mathbf{F}_i}{m} \text{ and } \frac{d\mathbf{r}_i}{dt} = \mathbf{v}_i \tag{32.1}$$

The force is obtained as the gradient of the **potential energy function** that describes the interactions between the atoms in the system:

$$\mathbf{F}_i = -\nabla V(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{32.2}$$

After the initial positions and velocities are chosen, Hamilton's equations of motion are numerically integrated to determine the position and velocity of each atom as a function of time. From the final positions and velocities of the atoms, the identity and kinetic energy of all ejected species can be calculated. Additionally, the atomic motions leading to the ejection of the molecule can be analyzed [Garrison et al., 2000; Garrison, 2001].

A limitation of MD is that atoms are assumed to obey the laws of classical mechanics. Therefore, quantum effects such as electronic excitation and ionization are not included. The questions that one is hoping to answer should be chosen so that quantum mechanical effects are not essential. Furthermore, it should be kept in mind how the assumption of classical behavior may affect the results [Garrison et al., 2000; Garrison, 2001].

The chemical interactions between atoms are modeled by potential energy functions that describe how the energy depends on their relative positions. Ideally, the potential energy function would be the solution to the electronic Schrödinger equation within the **Born–Oppenheimer approximation**. However, such a solution is not possible in simulations with thousands of atoms. Instead, a suitable mathematical function that models the physical behavior is used. Parameters in the equation are fit to available experimental data [Garrison et al., 2000; Garrison, 2001].

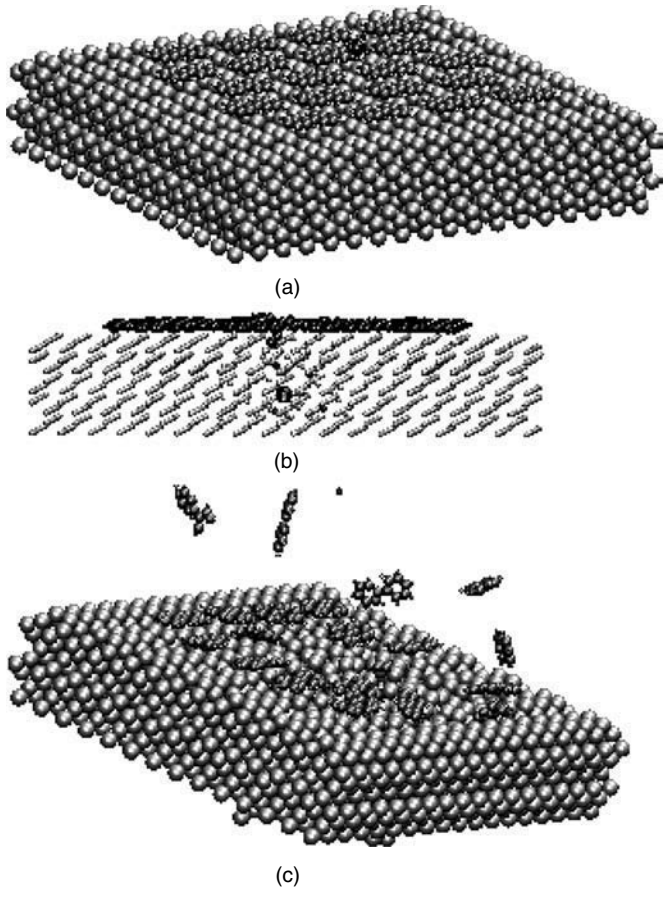


FIGURE 32.1 Collision cascade and ejection occurring for SF_5 bombardment of a monolayer of biphenyl molecules on Si{100}-(2x1). The Si atoms are represented by silver spheres, the carbon and hydrogen atoms are represented by dark gray spheres, and the S and F atoms are represented by black spheres. (a) An early snapshot in time that shows the SF_5 projectile as it moves toward the surface. (b) At 150 fs, the SF_5 projectile has penetrated into the open lattice and has broken up within the substrate. In this frame, the radii of the silicon spheres are drawn smaller so that the projectile can be seen within the substrate. (c) At 1500 fs, collision cascades initiated by the breakup of the projectile within the surface lead to the ejection of biphenyl molecules and molecular fragments.

Initially, the only available potential energy functions were **pair potentials** that depend on the distance between two atoms, such as the **Morse potential** and the **Lennard–Jones potential**. The total potential energy is calculated as a sum of the pairwise potential between each pair of atoms in the system. The **pairwise additive assumption** cannot take into account the effect of neighboring atoms. Therefore, this approach is not successful for modeling extended solids. Pairwise potentials are especially limited for chemistry applications, because they cannot accurately describe polyatomic molecules. For example, the pairwise additive approach would predict that H_3 is more stable than H_2 [Garrison, 2001]. The potential energy between each pair of atoms in a molecule depends on the coordination number of the central atom, the bond angle, and the torsion angle. In order to model chemical reactions, changes in hybridization must also be taken into account.

In the last decade, a great stride in these simulations was taken with the development of **many-body potentials** that include the effect of neighboring atoms on the interaction between a pair of atoms [Garrison et al., 2000; Garrison, 2001]. This development produced accurate potential energy functions for modeling face-centered cubic metals and silicon. Of particular interest to chemists is Brenner's reactive empirical bond-order (REBO) potential [Brenner, 1990], which varies the bond strength depending on

the coordination number, bond angles, and **conjugation** effects. This potential is able to model bond breaking and formation because atoms may change neighbors and their hybridization state. These sophisticated many-body potentials are blended with empirical pairwise potentials to model keV bombardment of organic films on metal surfaces [Garrison et al., 2000; Zaric et al., 1998; Townes et al., 1999; Nguyen et al., 2000; Garrison, 2001].

A limitation of the Brenner REBO potential is that it cannot describe long-range interactions between molecules. Recently, Stuart et al. have developed a reactive potential for hydrocarbons that includes both **covalent** bonds and intermolecular interactions [Stuart et al., 2000]. The adaptive intermolecular REBO (AIREBO) potential introduces nonbonded interactions through an adaptive treatment, which allows the reactivity of the REBO potential to be maintained. A possible problem with the introduction of intermolecular interactions is that the repulsive barrier between nonbonded atoms may prevent chemical reactions from taking place. The AIREBO potential corrects for this problem by modifying the strength of the intermolecular forces between pairs of atoms, depending on their local environment. For example, the interaction between two fully **saturated** methane molecules will be unmodified, producing a large barrier to reaction. The carbon atoms in two neighboring methyl **radicals**, on the other hand, will have a repulsive interaction that is diminished, or even completely absent, allowing them to react.

32.4.2 Applications of MD Simulations

MD simulations of the high-energy bombardment of organic films on atomic substrates with atomic and polyatomic projectiles have led to interesting insights about the mechanisms by which polyatomic projectiles enhance the ejection yield [Garrison et al., 2000; Zaric et al., 1998; Townes et al., 1999; Nguyen et al., 2000]. The simulations also have contributed information about the optimum conditions for the effectiveness of polyatomic projectiles. As a result of the simulations, three factors have been identified as important to the enhancement in yield with polyatomic projectiles:

Collaborating collision cascades
Open lattice structure of the substrate
Mass matching

First, molecules that have multiple contact points to the surface are ejected intact when several substrate atoms hit different parts of the molecule from below. Polyatomic projectiles enhance the emission yield by increasing the probability of adjacent collision cascades in the substrate, which can collaborate to lift the molecule gently from the surface, as shown in Figure 32.2.

Second, the nature of the substrate is a critical factor in the effectiveness of polyatomic projectiles [Townes et al., 1999; Nguyen et al., 2000]. The enhancement in yield is greater on a substrate with a more open lattice structure, such as silicon, than on a more closely packed substrate, such as copper. When SF_5 bombards an organic layer on copper, a densely packed solid, the polyatomic projectile breaks apart as it hits the surface and is reflected toward the vacuum. With the silicon substrate, on the other hand, the entire SF_5 projectile is able to penetrate the surface and break apart within the substrate, as shown in Figure 32.1. The breakup of the cluster within the lattice initiates upward-moving collision cascades that work together to lift intact molecules from the surface.

Third, polyatomic projectiles are most effective when there is mass matching between the atoms in the projectile and the atoms in the target solid [Townes et al., 1999; Nguyen et al., 2000]. When the mass of the atom (or atoms) of the projectile is much larger than the mass of the substrate atoms, the projectile passes through the solid without transferring much energy to the atoms in the top surface layers. When the projectile atom (or atoms) has much less mass than the substrate atoms, the projectile reflects from the surface, retaining much of its initial kinetic energy.

32.4.3 Concluding Comments on MD Simulations

The mechanistic insights obtained from MD simulations are significant for the further development of SIMS as an analytical tool. The simulations predict that the greatest enhancements with polyatomic

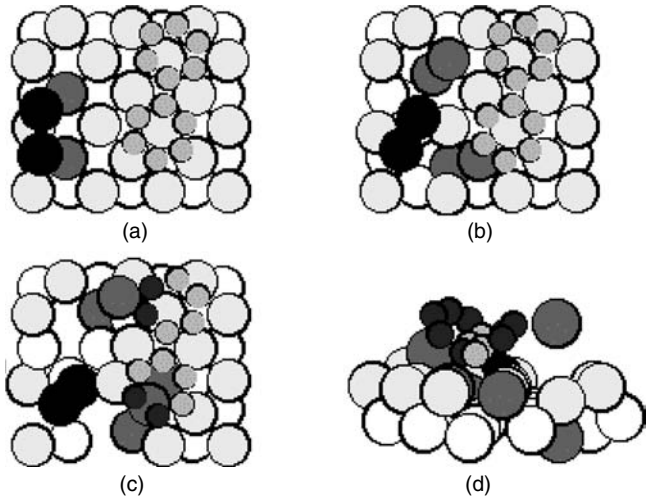


FIGURE 32.2 Schematic diagram illustrating the mechanism for the ejection of a biphenyl molecule with a diatomic projectile at 0.200 keV or 0.100 keV per atom. The incoming cluster atoms are black, and the biphenyl molecule is shaded gray. As atoms become part of the collision sequence leading to ejection of the molecule, they are shaded a darker gray. The two atoms in the dimer act collaboratively to initiate two adjacent collision cascades that lead to hitting carbon atoms in each ring of the molecule. (a) 45 fs, top view. (b) 63 fs, top view. (c) 84 fs, top view. (d) 104 fs, top view.

projectiles such as SF_5^+ and C_{60}^+ will be on organic solids, which have an open-lattice structure and are composed of light atoms. The development of focused polyatomic ion beams would have a significant impact on molecular specific imaging experiments, for which sensitivity is presently the limiting factor [Berry et al., 2001]. With the recent development of the AIREBO potential, the next challenge is to perform MD simulations of the high-energy bombardment of molecular solids, in which both short-range intramolecular forces and long-range intermolecular forces are present.

32.5 Modeling Organic Compounds

Organic chemistry encompasses all aspects of carbon-containing compounds, including the study of their chemical and physical properties, reactions, synthesis, and various uses. Organic chemistry is the foundation of biochemistry, polymer chemistry, materials science, medicinal chemistry, colorants, fragrances, the pulp and paper industry, and many other areas. The history of organic chemistry has been relatively brief: its formal beginnings were in 1826, with Friedrich Wöhler's preparation of urea that debunked the hypothesis of a **vital force**. Since that time, the use of models has been fundamental to the field. Organic chemistry is driven largely by the modeling tools that organic chemists use. Computational chemistry is currently one of the most powerful, and most rapidly developing, of those tools.

Two key models for chemical bonding guide the work of organic chemists: **valence bond theory** and molecular orbital theory. Organic computational chemistry is founded in both theories. Organic chemists use computational chemistry as a tool to solve three general types of problems:

Predicting or rationalizing chemical or physical properties
Predicting or rationalizing the stabilities of compounds
Predicting or rationalizing the reactions of compounds

The computational software used by particular organic chemists depends on personal preferences and abilities, and it varies in robustness and hardware availability. There are many sources — some free — for computational chemistry software. Because not all software packages use exactly the same programs or parameters, the publisher of the software and the specific program(s) used for each research project must

be stated in any publications on that project. Generally, the hardware is less of a significant limiting factor (although the hardware can determine the speed of the calculation) and often is not mentioned in the publication.

The front end of most computational chemistry software is now so automated and polished that its effective use in computational chemistry is as simple as the following procedure:

Draw a 2-D structure of a molecule
Select a computational method from a drop-down menu of listed options (which can include **empirical**, **semiempirical**, and **DFT/*ab initio*** methods)
Wait for the computer to generate an output file, consisting of a three-dimensional structural representation, maps of potential energy surfaces, and a listing of computed values
Look over the output to see what it reveals

The computational time depends on both the computer and the complexity of the computation. For example, on a robust desktop PC, a moderately sized organic compound with 20 to 40 atoms might take less than a minute to compute using an empirical method, several minutes to an hour with a semiempirical method, and a week or more for a DFT/*ab initio* process. Unless the chemist is a specialist in computation, most of the personnel time is spent on the correct selection of the computational method to be performed (definition of the model), and then on the analysis of the output. End users are indebted to those computational chemists who have pioneered these techniques and made the tools of computational chemistry available to all chemists. For example, Professor Norman L. Allinger at the University of Georgia led a group that developed MM3, one of the more popular methods currently in use [Allinger et al., 1989].

32.5.1 Empirical Solutions

For many organic chemists, empirical methods, known collectively as **molecular mechanics** (MM) routines or *force-field methods*, are the easiest, fastest, and most generally used computational tools. This is not surprising, because these programs were originally designed for application to organic compounds. Conceptually, these methods fit easily into the historical development of valence bonding theory, and they are rooted mathematically in expressions of Hooke's law. By summing the energies of all bonded atoms, an estimate of the relative energy of the entire structure is derived. The force-field methods are, in a sense, mathematical extensions of valence bond theory's physical constructs of bonds and molecules, which contributed to early discoveries such as the alpha helix and **conformations** of cyclic and acyclic structures.

In practice, the proper selection of both the force-field routine and the parameter sets used for the class of compound being evaluated is crucial to getting a reliable result. Most bench chemists use an unmodified parameter set provided with the software. An empirical program will typically evolve over several years. The identities of these programs may be designated by year (e.g., MM3[92] or MM3[96]), by special characters (e.g., MM2*), or by a generic description of change (e.g., MM2 augmented). A problem with any computational result, but especially with an empirical calculation, is that the result occurs even if it has no practical validity. Thus, the user must know the limitations of the method and verify the reliability of any calculation.

Despite this limitation, the development of force-field programs over the years has created an ability to model reliably many classes of organic compounds, making these programs the first choice of many chemists. The basic set of equations used to model an organic compound computationally partitions and sums the contribution of the several factors to total energy. Those factors are related to the mass of the atoms in the molecule, the known preferences for bond angles and bond lengths between those atoms, **van der Waals forces** between atoms, interaction of bonds with dihedral relationships, and electrostatic interactions between atoms. In one commercially available software package, CAChe, the MM3 augmented routine includes bond stretch, bond angle, dihedral angle, improper **torsion**, torsion stretch, bend bend, van der Waals forces, electrostatics, and hydrogen bond terms. In comparison, the augmented MM2 routine in the same package does not include the bend bend interaction.

Parameters developed to describe a term in one force field are not necessarily useful in other force fields to describe the same term. It takes significant effort to develop and validate new parameters when a new routine is developed to solve a new problem or when an existing force field is modified to deal more effectively with a new functional group [Woods et al., 1992; Todebush et al., 2002]. In most MM programs, the molecule is modeled without molecular solvent or other interacting molecules, effectively modeling the preferences of a compound as a gas. Because organic chemistry is often performed in the solid state or solution phase, the influences of these states must be considered separately. As previously mentioned, the major limitation of MM methods is that they do not describe electronic properties of molecules.

Case Study

To illustrate the use of computational tools, we review the application of molecular mechanics in the development of the synthesis of a vasopressin receptor antagonist, SR 121463 A, as presented in [Venkatesan et al., 2001].

A key step in this synthesis involves the stereoselective reduction of a ketone, which results in the formation of two different alcohol products, designated **syn** and **anti** isomers. (The *syn/anti* nomenclature describes the relative position of the alcohol to the amide carbonyl in the product.) While reduction with a commonly employed reagent ($LiAlH_4$) gives acceptable initial results (4:1 syn:anti), Venkatesan et al. sought improved selectivity. They wanted to avoid loss of product by minimizing the amount of the anti product formed. They also wanted to reduce the effort needed to separate the syn from the anti alcohol.

The researchers used MM routines in MacroModel, a software package that allows for incorporation of solvent effects by means of a continuum model. They demonstrated computationally that, in the preferred conformation for the syn alcohol in solution, the alcohol was in an **equatorial** position. (This position is preferred over an alternative structure with the alcohol in the **axial** position by 2.0 kcal/mol using MM3*, by 1.1 kcal/mol using **MMFF** and, as determined for comparative purposes here, by 1.5 kcal/mol using MM3 in CAChe without solvent parameterization). This was in agreement with known general stabilities for equatorial and axial alcohols. However, upon examination of a solid state X-ray structure of the syn alcohol, it was clear that the alcohol in the syn compound was axial. The researchers rationalized that the rather small energy preference for the equatorial alcohol in the syn compound was easily outweighed in the solid state by the increased stabilization from intermolecular hydrogen bonding in the solid state when the alcohol was axial.

To explain the increased preference for production of the syn alcohol when using lithium cation-derived reagents (as opposed to sodium cation reagents, which gave syn:anti product ratios of only 3:1), the starting material was modeled when complexed with a cationic replacement (ammonium ion) for the lithium cation (because the parameters for Li atom were not in the programs used). If the cation is coordinated to both the ketone and the amide carbonyls, the modeled minimum energy structure adopts a **twist boat conformation**, which is only 2.7 kcal/mol higher in energy than the normally expected **chair** structure. If the compound were to react exclusively from the di-coordinated twist boat structure, then the expected product would be the syn alcohol. They also determined that the barrier for conversion from the chair to the twist boat (which requires adoption of a higher energy structure, known as a **half chair** conformer) was only 4.0 kcal/mol. This represents a significant increase in the stability of the twist boat structure compared to the normally more stable chair structure. In the absence of other influences, the chair form is generally 5.5 kcal/mol more stable and the barrier for interconversion (via the half chair) is about 10 kcal/mol. While this does not entirely account for the increased preference for forming the syn alcohol as one varies the reducing reagent from $NaBH_4$ to $LiAlH_4$ to L-selectride (which gave a 66:1 syn:anti preference), it aids in understanding the process, which is important if this process of stereoselective reduction is to be extended to use in other systems.

32.5.2 Semiempirical Methods

Organic chemists often use more complex methods than empirical force-field methods in order to accurately predict chemical process, especially if the nature of the electronic interactions is a controlling factor. In computational organic chemistry, molecular orbital theory provides an effective methodology. On a simple level, many problems can be evaluated by taking into consideration only the electrons most intimately involved in a chemical process. Semiempirical methods do this, by applying approximations based on empirically derived data to the **Hamiltonian** equations used to model the compound.

For example, an estimate of electronic transitions in ultraviolet-visible spectrophotometric measurements is modeled and understood by looking at the **frontier molecular orbitals** (FMO) of the **pi** system undergoing a transition, rather than including all electrons in the molecule (nonbonded and **sigma** bonded). A semiempirical method, such as **ZINDO**, is often used for this purpose. Other semiempirical methods make different approximations to simplify the computational task, and their approximations are included as look-up parameters from experimentally determined data. Three that are commonly found in software packages are the modified neglect of diatomic overlap (MNDO), the Austin models (AM1, AM3), and the third parametization of the MNDO model (PM3). The best strategy for selecting a semiempirical method is often simply to use the one that comes closest to fitting the experimental truth.

Although FMO had been used qualitatively for many years by organic chemists to rationalize chemical reactions [Fleming, 1976], the semiempirical methods, which may or may not provide accurate quantitative results, often fail even at a qualitative level.

32.5.3 *Ab Initio* Methods

The most robust computational methods employ full quantum mechanical methods. Even here, approximations are made, but they are minor compared to semiempirical methods. Such methods are often referred to as *ab initio* calculations, because they derive the energy of the molecule entirely from its native collected electrons.

There are different ways to perform *ab initio* calculations, allowing the electrons to express differing degrees of sophistication and freedom. It is possible to include all electrons, or to consider only the valence electrons, treating the nonvalence electrons as the so-called *frozen core*. Borden and Davidson elucidate the need to include all electrons with full electron correlation in complex computational chemistry problems [Borden and Davidson, 1996]. Although they take more time and employ more sophisticated calculations based on application of the Schrödinger equation, these *ab initio* methods (and DFT-type *ab initio* calculations) are required to evaluate **transition states** of **pericyclic processes**, chemistry of the **excited state**, and studies on radical structures or processes that have potential radical intermediates. Problems that must be addressed by DFT/*ab initio* methods are generally evaluated in stages: using empirical methods to get an initial structure, using semiempirical methods to get a refined structure, then crunching out the DFT/*ab initio* calculations.

Case Study

Here, we summarize the use of DFT methods to ascribe the stable conformers of a set of similar molecules, each of which undergoes an intramolecular **Diels–Alder** reaction, and then to ascertain the transition states of the Diels–Alder reactions [Bur et al., 2002].

Bur et al. set out to determine why four seemingly similar molecules required such vastly different experimental conditions to undergo an intramolecular Diels–Alder reaction. In the first part of the analysis, they identified a problem in the use of empirical methods (MM2* or MMFF using MacroModel7.0) for determining of the lowest energy conformation of the starting materials. They employed Monte Carlo searching in conjunction with applied force fields to identify the lowest energy structures. However, the lowest energy output placed two hydrogen atoms too close to each other in the molecule (in one case at a distance of only 2.40 Å, which implies contact between the two).

To resolve this issue, the researchers employed DFT calculations with **B3LYP/**3-21G* **basis sets** using Gaussian 98, followed by further refinement using 6-31G* to obtain stable conformers that had reasonable interatomic distances. They then used a 6-31G* basis set to model the Diels–Alder transition states and found that the results were qualitatively useful (the compounds requiring external heat in excess of 100°C to bring about a reaction also had computed activation energies of about 4.7 kcal/mol over the reactions known to occur at room temperature). However, the authors recognized that the results were not quantitative. By comparing the calculated activation energies for the two slower reactions (requiring heating to 145°C or 110°C, respectively), they saw that the energy difference of the transition states, only 0.3 kcal/mol, could not explain the temperature differences needed to bring about reaction.

Recognizing that the difference in transition state energy alone would not account for the observed discrepancy in the reactivities of the systems, they reexamined the conformational profiles of the starting materials for other factors. They identified two computationally modeled factors for the conformations of the reactants that corresponded to the reaction rates. The first was the influence of the carbonyl linkage (present only in the faster-reacting compound) between the 2-pi electron system and the 4-pi electron system, which resulted, as one might expect, in a more favorable alignment in the structure. The second accelerating feature was a preferential rotation about the C–N bond in the linkage that again favored a conformation of prealignment between the 2-pi and 4-pi systems in the faster-reacting materials.

Computational analysis is now central to the practice of organic chemistry (as spectroscopic analysis had become in the middle of the 20th century). The importance of computational chemistry for organic chemists is likely to increase as they find more useful tools for predicting and rationalizing chemical processes — and as computational chemists continue to advance and refine their science.

32.6 Computational Organometallic and Inorganic Chemistry

The application of modern computational techniques to inorganic and organometallic chemistry has truly undergone a renaissance during the past decade. Stoichiometric and catalytic metal reactions have attracted great interest for their many applications in industrial and synthetic processes. Metal reactions are critical in many thermodynamically feasible processes, because they accelerate the reaction by opening a lower **activation energy** pathway. These metal-centered reactions consist of one or more elementary reactions, such as substitution, oxidative addition, reductive elimination, migratory insertion, hydrogen exchange, β-hydrogen transfer, σ-bond **metathesis**, and **nucleophilic** addition.

However, unlike organic compounds, inorganic compounds are unsuited to the application of empirical molecular mechanics calculations. As mentioned in the case study in Section 32.4.1, parameters for many of the most common elements (e.g., Li) are not included in these programs, and the terms used were not developed to handle issues relevant to inorganic compounds (e.g., expanded octets, multiple stable valences, etc.).

Recent progress in computational chemistry has shown that many important chemical and physical properties of the species involved in these reactions can be predicted. Calculated values, such as predicted geometries, **vibrational frequencies**, bond dissociation energies, and other chemically important properties, have become reliable enough to complement and sometimes even challenge experimental data, especially in those cases in which experimental results are difficult to obtain. The most challenging aspect of this area has undoubtedly been to model the **metallic** elements reliably and efficiently. Metals, particularly the ***d***- and ***f*-block elements**, typically present three main challenges for modeling:

The large number of orbitals, many of which are core orbitals
The **electron correlation** problem, which is accentuated by the presence of low-energy excited states
Relativistic effects for the heaviest metals

32.6.1 Semiempirical Methods

A range of quantum chemical methodologies can be used to study inorganic and organometallic compounds. The semiempirical quantum mechanics methods have great latitude in the number and type of approximations made to the full Schrödinger equation that involve the replacement of quantities that are difficult to determine with experimental or theoretical estimates or the removal of interactions, like electron interactions, which are thought to be of lesser importance. The trade-off for such approximate methods is accuracy vs. computational efficiency. For many inorganic or organometallic materials, the sacrifice of accuracy for speed is problematic because of the challenges listed previously.

The use or extension of approximate semiempirical methods necessitates a parameterization phase. Here, it is necessary to determine those parameters that maintain computational efficiency, while maximizing the model's descriptive and predictive power. Ideally, the parameterization process should incorporate the full range of motifs that characterize a chemical family. Therefore, one major issue for parameterizing metal-containing compounds is the development of a robust parameterization to handle a diverse set of influences. Such chemical diversity may be defined as the ability of metals to stabilize distinct bonding environments involving different bond types, bound-atom types, **spin** and **formal oxidation** states, **coordination numbers**, and geometries. For these reasons, it is difficult to use semiempirical methods when the calculations involve metal atoms.

32.6.2 *Ab Initio* Methods

At the other end of quantum chemical methodology spectrum lie *ab initio* methods. These techniques employ computations derived from theoretical principles without inclusion of experimental data. While *ab initio* methods such as **Hartree–Fock** (HF) and **Møller–Plesset** (MP2) have been utilized extensively for organic compounds, their application to metal-containing systems is limited. Results from such calculations involving metals have been more-or-less useless: bond lengths were wrong by tenths of angstroms, the relative energies of **isomers** were often wrong, and the relative energies of possible spin multiplicities were wrong, to name a few instances. For organometallic compounds, the primary problem with application of *ab initio* methods was founded in the change in the error with bond length that exceeds the change in the actual energy. This swamped other errors that result from use of simplified models for complicated **ligands**, neglect of the solvent, and neglect of relativistic corrections.

However, two techniques have been developed to deal with these challenges: quasi-relativistic **effective core potentials** (ECP) (also commonly referred to as *pseudopotentials*) and **density functional theory** (DFT). There is general agreement that DFT methods are superior to classical *ab initio* methods at the HF and MP2 levels for the calculation of metal compounds. This is because the accuracy of the DFT results is similar to, or in some cases even better than, MP2 data, and the computational time costs are significantly less. DFT combined with ECPs are now considered the standard operating procedure for handling inorganic and organometallic compounds. Indeed, computational inorganic and organometallic chemistry is almost synonymous with DFT for medium-sized molecules. This is due in part to computational improvements but also, perhaps more importantly, to the inclusion of these techniques in powerful, yet relatively user-friendly computational chemistry packages, such as Gaussian, Spartan, Jaguar, GAMESS, MOLPRO, CAChe, Hyperchem and ADF.

Powerful DFT methods incorporating ECPs allow many topics of interest to inorganic and organometallic chemists to be studied. A significant body of literature has been produced concerning theoretical studies of **transition metal**-catalyzed chemical reactions including industrially relevant processes, such as Ziegler–Natta polymerization, copolymerization, hydroformylation, and the water-gas shift reaction [Torrent et al., 2000]. Other catalytic systems studied include hydrogenation, epoxidation, dihydroxylation, and thioboration [Torrent et al., 2000]. It remains to be seen whether there is a catalytic system unamenable to computational study, such as **heterogeneous** catalysts, for which it is difficult to characterize intermediates experimentally.

Another topic of interest to inorganic and organometallic chemists is the nature of bonding. Included in bonding studies is the multiple bonding to ligands and other metals, both transition and main group

[Frenking and Fröhlich, 2000; Cundari, 2000]. The aim of most theoretical investigations of the chemical bond is to find a correlation between the chemical behavior of a molecule or its physical observables and calculated data, such as charge distribution or orbital structure. When including metals in theoretical models, a considerable problem is understanding the energetic contributions to the chemical bonds in terms of what the electrons are doing. Much progress has been made in understanding the nature of bonding to metals, and the future of this area is robust with possibilities.

The advent of faster computers and better algorithms has made possible new areas of theoretical work with metals. Until recently, there has been a paucity of work with **actinide** complexes, undoubtedly due to the experimental and computational difficulties in handling these elements. In addition to the problems mentioned earlier, the application of theoretical electronic structure methods to actinide complexes has long been deterred by several well known challenges. The ability to accommodate f orbitals with computational efficiency; the incorporation of a larger number of valence electrons; dynamic electron correlation effects; the large number of low-lying, near-degenerate states; the severe relativistic effects caused by the large atomic numbers of the actinides; and the overall large size of actinide complexes place extremely high demands on the choice of suitable basis sets, efficient numerical algorithms, and computational resources.

With DFT methods, actinide complexes are no longer ignored [Li and Bursten, 2001]. DFT methodology allows experimentally important properties (such as the geometry, vibrational frequencies, and infrared absorption intensities) to be calculated, even for large actinide systems. Many aspects of actinide chemistry are experimentally challenging, so reliable theoretical calculations provide a valuable adjunct to experimental studies and can provide theoretical interpretations of experimental results.

Accurate quantum chemical treatment of transition metal complexes in biochemical systems is a relatively new area [Siegbahn and Blomberg, 2000]. With each passing discovery of an important metalloenzyme or metal-mediated biological process, the ability to predict *in vivo* function of metals in biological systems is becoming increasingly important. There are two striking differences in the application of computational chemistry to biological systems, in contrast with other chemical systems such as **homogeneous** catalysis. The first difference is that the overall chemistry that affects biochemical complexes is considerably more complicated than in a typical catalytic cycle of a laboratory process. The second difference is that the amount of experimental information on biochemical systems is immense. Often, decades of experimental investigations are done by researchers focusing on only one system. To make significant contributions, the large amount of biochemical information must be put in the context of a quantum chemical modeling, which is not a trivial matter.

As with inorganic and organometallic systems, metal-containing biological systems must be studied with methods that incorporate correlation (e.g., DFT), because significant errors occur in **thermodynamic** properties, which are at least an order of magnitude larger. It is safe to predict that almost all future studies of biological transition metal systems will be accomplished using more accurate methods, such as gradient-corrected DFT. Systems that have been treated quantum-mechanically include accurate geometric determinations of the metal binding site in blue copper proteins, mechanisms of methane monooxygenase, and water oxidation in photosystem II. The number of applications is still not large, but this area will grow rapidly in the future. It is increasingly common in the computational study of large biological systems to use hyphenated methodologies, such as QM-MM. Quantum mechanical (QM) methods are employed to study the inner working of the active site or the metal center of the metalloprotein (where electron correlations are of paramount importance); empirical methods (MM) are used to define the carbon skeleton scaffolding that comprises the bulk of the system. The interface of the quantum mechanical region and the empirical region of the system is difficult, and discussion of how to address that problem is treated elsewhere [Carloni et al., 2002; Monard and Merz, 1999].

Finally, while recent work has focused on the use of DFT, it must be noted that such methods do not solve all problems. The magnetism and spin states of multimetal clusters and solids are not well determined by DFT; these are most easily described by an empirical Heisenberg Hamiltonian. Detailed potential curves and states of very small molecules to spectroscopic accuracy require configuration interaction (CI) methods. Solvent effects on spectra remain a difficult problem as well. While DFT certainly has revolutionized inorganic and organometallic computational chemistry, it is not a panacea for all problems

facing the discipline. However, the growth and development of computational inorganic chemistry has been unprecedented over the past decade and will likely continue to expand in coming years.

32.7 Use of *Ab Initio* Methods in Spectroscopic Analysis

It is now possible to carry out molecular orbital *ab initio* calculations on a reasonably complex molecular structure. These calculations can yield details on a number of important molecular properties, such as the following:

Geometry in ground and excited states
Atomic charge and dipole moment
Molecular energy
Conformational stability and structure of macromolecules and biomolecules
Vibrational frequencies
Infrared intensities
Raman activities
Electrostatic potential energy
Force constants
Dynamics of molecular collision
Rate constants of elementary reaction
Simulation of molecular motions
Thermodynamic properties

The challenge here is determining how to obtain this information and what tools are needed to do so. As previously noted, much *ab initio* computational chemistry is based on the Schrödinger equation, $H\Psi = E\Psi$, developed in 1926 by the Austrian mathematician and physicist Erwin Schrödinger. This is a single equation, whose solution is the wave function for the system under consideration and describes the spatial motion of all particles of the molecular system. In order for the equation to work, the wave function must satisfy certain properties. Unfortunately, the exact solution of this equation can be used to calculate the energy of only a single atom: the one-electron hydrogen atom. Exact solutions are not possible for even the two-electron helium atom, or for any other elements or compounds. Still, the Schrödinger equation is utilized for larger systems consisting of many interacting electrons and nuclei. To do this, a number of mathematical methods that use approximation techniques such as the **variational** theorem, self-consistent field theory (Hartree–Fock approximation), and **linear combination of atomic orbitals** (LCAO) are applied to solve this equation and to describe the atomic and molecular structures.

32.7.1 Hartree–Fock Approximation

The central difficulty in solving the Schrödinger equation is dealing with electron–electron interactions. The energy of a particular electron depends on the electric field generated by the nuclei of the elements and all the other electrons in the system. According to the **variational principle**, we can approach an accurate solution of many electron wave functions once a set of adjustable coefficients is used with each orbital to minimize the total energy of the system. All the contributing one-electron functions are then varied until the energy obtained is at its lowest value. Such orbitals are referred to as *self-consistent field* or **Hartree–Fock** molecular orbitals.

The advantage of this method is that it allows us to interpret the molecular properties as derived from those of the constituent atoms. Further, this method assumes that the electrons are moving in an environment that is an average potential of the other electrons. Thus, the instantaneous position of an electron is not influenced by the presence of neighboring electrons. It should be noted that the HF approach fails to take into consideration the Pauli exclusion principle that two electrons cannot have the same quantum numbers, so HF does not account for electron–electron repulsion. Several methods are used to represent this electron correlation, as mentioned in the following section.

32.7.2 Electron Correlations

The correlation of electrons is crucial in studying the optimization of structural parameters, energies of conformational stabilities, and vibrational frequencies, each of which is essential to spectroscopic analysis. In HF calculations, as mentioned earlier, every electron is affected by the average of the other electrons in the atoms but is insensitive to any individual electron–electron interaction. Several electron correlation methods are used, such as Møller–Plesset to the *n*th order of correlations (MPn), configuration interaction (CI), multiconfigurational self-consistent field (MCSCF), generalized valence bond theory (GVB), and coupled cluster theory (CC). Including correlation functions in the calculations results in more accurate computational energies and structural parameters.

32.7.3 Gaussian Basis Functions

In 1930, Slater defined a particular set of functions associated with the molecular configuration. This set, which depends only on the nuclear charge, results in what are known as Slater-type atomic orbitals (STO), which have exponential radial components represented as $e^{-\xi r}$. Slater functions are of limited use because such integrals must be solved by numerical methods and are not well suited to numerical calculations in molecular systems. In 1950, a suggestion was made by S.F. Boys that the atomic orbitals should be expressed in terms of Gaussian-type orbitals (GTOs), in which the exponential radial parts are represented as $e^{-\alpha r^2}$, instead of $e^{-\zeta r}$ as in Slater-type atomic orbitals. The exponential terms ξ and α are constants that determine the size of the orbital.

The advantage to using GTOs is that they do not require numerical integration, and the product of two Gaussian functions is another Gaussian function. The disadvantage to using Gaussians is that the atomic orbital is not well represented by a simple Gaussian function but by a sum of several Gaussian functions. That is, a Gaussian function decays too slowly and thus lacks a cusp at the origin, as is required by STO representations. (See Figure 32.3). Although fast, modern computers make the use of multiple Gaussian functions feasible, there is still some debate over the intrinsic value of using multiple Gaussians. One school of thought holds that increasing the number of basis functions will improve the model for the calculation of structural parameters. Others believe, however, that this will not improve the model but give rise to erroneous results.

In any event, the basis functions for multiple Gaussians are further modified by adding polarization and diffusion functions for hydrogen and heavy atoms. The choice of basis set affects the computation time to perform the calculation to the order N^4, where N is the number of basis functions. The smallest basis set used in the calculations is called the *minimal basis set*, for example, STO-3G. Despite this cost in computational time, polarization functions are still used because they often produce more accurately computed geometries and frequencies.

32.7.4 Notation

In the literature, there are two popular types of basis sets: designated STO-nG and the Pople or split valence set. The latter is designated a-bcG, where n, a, b, and c correspond to number of Gaussian functions used to form each Slater-type orbital. In STO-3G set, the 3 corresponds to the use of three Gaussian functions to mimic Slater orbitals; STO-4G indicates that four Gaussian functions are used, and so on. In split valence sets, such as 3-21G (which is the minimal basis set in this family), three primitive Gaussian functions are used to describe the core shell (nonvalence orbital). The valence shell is described by the linear combination of two primitive Gaussian functions and one primitive Gaussian function, so that a total of six GTOs are utilized to mimic the Slater-type orbital. In 6-31G, the core orbital is represented by six primitive Gaussian functions, and the valence shell is described by two functions that are linearly combined: one with three primitive Gaussian functions and the other with one.

As stated, these two basis sets (3-21G and 6-31G) do not allow for the polarization of the orbitals. This means that the electrons are not allowed to occupy orbitals other than those they would occupy based

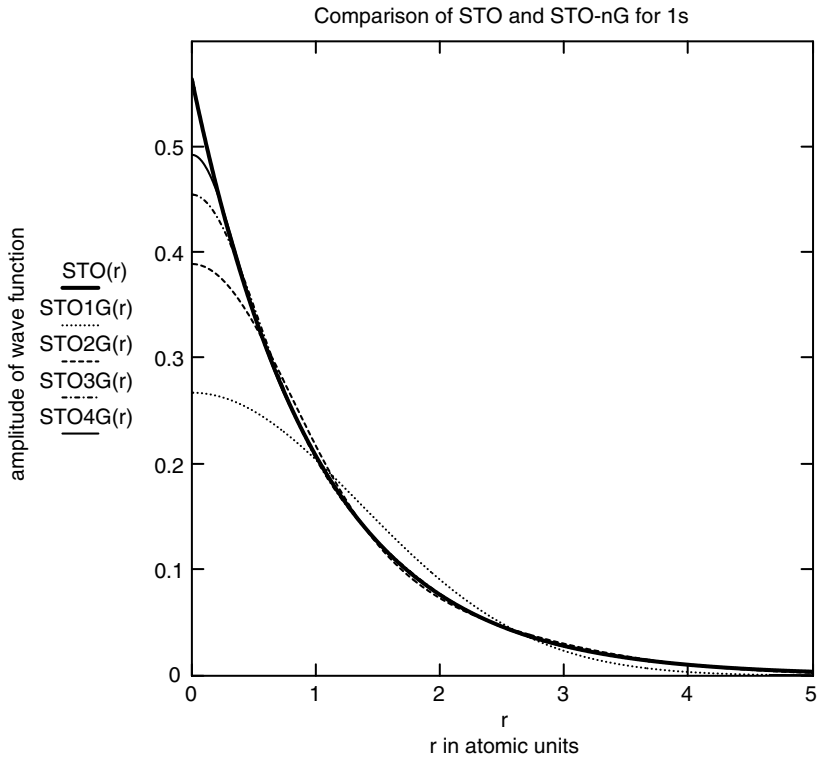


FIGURE 32.3 Representation of Slater-type orbital (heavy solid line) using various Gaussian-type orbitals.

on standard electron configuration principles. Polarization allows orbitals to change shape (e.g., from a spherically symmetrical *s* shape to a dumbbell *p* shape), and this is an important feature of orbitals that can be achieved by applying a polarization function to the basis set. This is designated by adding one or two asterisks to the basis set: for example, 6-31G* or 6-31G**. The polarization functions give the **wave function** flexibility to change its shape by adding another set of primitives. Polarization functions are crucial in the computation of structural parameters and vibrational frequencies. In the first case (6-31G* or 3-21G*), the single asterisk means that the basis set has a polarization function on the heavy atoms or nonhydrogen atoms in the molecule. Addition of a second asterisk (6-31G** or 3-21G**) indicates the use of a polarization function for the hydrogen atoms, so that polarization of all orbitals is manifest. A diffusion function can also be added to the basis set, indicated by a plus sign (as in 6-31^{+}G and 6-31^{++}G). This is important for calculations involving **anions**. A single plus sign indicates that the diffusion function has been added to the heavy atoms (nonhydrogen atoms); adding two plus signs indicates that the diffusion function is added to all atoms.

32.7.5 Vibrational State and Spectra

In order to carry out the quantum mechanical treatment of molecular vibration, it is necessary to introduce a new set of coordinates, $Q_k, k = 1, 2, 3 \ldots 3N$, called *normal coordinates*. Each normal mode of vibration can be characterized by a single normal coordinate Q, which varies periodically. A given normal coordinate is a measure of the amplitude of a specific normal mode of vibration. The energy levels of a harmonic oscillator are given by the expression $W = (v + \frac{1}{2})h\nu$, where v is the quantum number, ν is the classical frequency of the system, and h is Planck's constant. Consequently, the vibrational energy of the molecule

with several classical frequencies ν_k is described as shown in Equation 32.3.

$$W = \left(\nu_1 + \tfrac{1}{2}\right)h\nu_1 + \left(\nu_2 + \tfrac{1}{2}\right)h\nu_2 + \cdots + \left(\nu_{3N-6} + \tfrac{1}{2}\right)h\nu_{3N-6} \tag{32.3}$$

In other words, every frequency coordinate Q_k has associated with it a quantum number ν_k and a normal frequency ν_k, which is the classical normal frequency of vibration.

In order to obtain a more complete description of the molecular motions involved in the normal modes of a molecule, a normal coordinate analysis is performed. The force field in Cartesian coordinates can be obtained by the Gaussian program from the calculations of Hartree–Fock or Møller–Plesset perturbation level of theory utilizing any basis set. The force constants that result from the *ab initio* calculations are then used to obtain the vibrational frequencies for infrared intensities and Raman activities. Initially, a scaling factor of 1.0 is applied to produce the pure *ab initio* calculated frequencies, called *unscaled frequencies*. Since they are the result of *ab initio* calculations, the predicted frequency is always higher than the observed frequency (in accordance with the variational principle). To compensate, scaling factors of 0.88 for the CH stretches, 0.9 for the CH bends and heavy atom stretches, and 1.0 for all the other modes, such as torsional modes (rotation about single bonds), are used to calculate the scaled frequencies and the potential energy distributions.

The calculated Raman spectra are simulated from the *ab initio* calculations to generate scaled predicted frequencies and Raman scattering activities. The Raman scattering cross section, $\partial_j/\partial\Omega$, which is proportional to the Raman intensity, can be calculated from the scattering activities and the predicted frequencies for each normal mode [Amos, 1986; Polavarapu, 1990]. To obtain the polarized Raman scattering cross sections, the polarizabilities are incorporated into S_j by $S_j\ [(1-\rho_j)/(1+\rho_j)]$, where ρ_j is the depolarization ratio of the jth normal mode. The Raman scattering cross sections and the calculated scaled frequencies are used with a Lorenzian function to obtain the calculated spectrum. Infrared intensities are calculated based on the dipole moment derivatives with respect to the Cartesian coordinates. The derivatives are taken from the *ab initio* calculations and transformed to normal coordinates by

$$\left(\frac{\partial\mu_\mu}{\partial Q_i}\right) = \sum\nolimits_j \left(\frac{\partial\mu_\mu}{\partial X_j}\right) L_{ji} \tag{32.4}$$

where Q_i is the ith normal coordinate, X_j is the jth Cartesian displacement coordinate, and L_{ji} is the transformation matrix between the Cartesian displacement coordinates and the normal coordinates.

The infrared intensities are then calculated by

$$I_i = \frac{N\pi}{3c^2}\left[\left(\frac{\partial\mu_x}{\partial Q_i}\right)^2 + \left(\frac{\partial\mu_y}{\partial Q_i}\right)^2 + \left(\frac{\partial\mu_z}{\partial Q_i}\right)^2\right] \tag{32.5}$$

The literature contains several representative examples of predicted Raman and infrared spectra derived from representative compounds [Guirgis et al., 2002; Guirgis et al., 2001; Mohamed et al., 1999].

32.8 Research Issues and Summary

In many ways, computational chemistry is just beginning to show its potential as a tool for explaining reaction processes. However, the greatest potential applications for computational chemistry seem likely to come from predictions for probable outcomes in complex chemical systems. Experimental chemistry will always be important, since every prediction must be confirmed experimentally. However, theoretical predictions already serve to guide experiment, and computational chemistry raises the power of that predictive ability to a new level.

As an example of the future predictive power of computational chemistry, consider the evidence presented in Section 32.3.3 and Section 32.5.2.1 of this chapter. As a further example of the promise for the predictive power of computational chemistry and the increasingly important role that computational chemistry is likely to play in directing experimental work, we close with the following case study.

Case Study

This case study involves stereocartography [Lipkowitz et al., 2002]. The issue of **chiral** control in chemical processes is of enormous importance. The FDA now requires biological testing of both **enantiomers** of chiral agents. The synthesis of a chiral material adds significant cost and difficulty to any preparative procedure. For these reasons and others, the ability to control the chirality of a chemical process is crucial. Moreover, the ability to achieve that control in a catalytic manner has obvious financial implications to the chemical industry. A major limitation in catalytic processes is obtaining a fundamental understanding of how the catalyst achieves its effect of accelerating a process, and, in the case of a chiral reaction, how the catalyst differentiates (and thus accelerates) the rate of one chiral pathway from another.

Many catalysts are metal-based systems. From a computational standpoint, this makes the modeling situation very complex. The working premise of this particular protocol was the readily believable but previously unproven hypothesis that an effective catalysis in a chiral system is one in which the chirality of the catalyst is as close as possible to the reactive site of the catalyst. With this assumption, a computational paradigm was needed to evaluate the effectiveness of known chiral catalysts. This paradigm maps the "stereodiscriminating regions around a chiral catalyst — hence the term stereocartography."

The first step was to make the center of mass of the catalyst at 0,0,0 on the Cartesian coordinate system, surrounded by a uniform three-dimensional grid. The second step was to put a transition state structure at grid points and compute the intermolecular energy (between transition state and catalyst) using molecular mechanics. The transition state–catalyst interaction was modeled 1,728 times at each grid point, using different alignments each time. For a typical analysis, this resulted in 950 million calculations! This is clearly a situation that calls for use of a highly parallel computational configuration, known as a *Beowulf cluster*.

To perform the analysis, parallel computing was employed "using a loosely networked cluster of SGI machines and on a small cluster of 26 AMD Athlon processors running a Linux operating system." The software used for semiempirical calculations of the transition state structures was Spartan 5.0, available from Wavefunction, Inc., which uses PM3 with a transition metal parameter set. **AMBER**, in MacroModel 7.0, was used for MM calculations.

The catalyst structures were obtained from the Cambridge Structural Database (Wavefunction), with PM3 optimization when necessary (i.e., when the database structure was not identical to the catalyst). To achieve chiral discrimination, the **R** and **S** transition states both were modeled within the grid, and the difference map (by subtraction) of the **electrostatic potentials** of the two **diastereomeric** complexes was examined. In the difference map, stereodiscrimination was revealed by the presence of electrostatic potential — if that potential were close to the site of the chemical transformation, the hypothesis would be confirmed.

Using this method, the authors examined 18 catalysts that were well described experimentally. Strikingly, 17 of the 18 known catalysts were shown to obey the Lipkowitz hypothesis. This level of success will assuredly lead many experimenters to model their catalysts with the Lipkowitz method before going to the difficulty of synthesis and experimental evaluation of their effectiveness. The authors appear to have made great progress toward their goal of quantifying the factors that influence the chiral induction of the catalytic system "so that ligands for use as **asymmetric** catalysts could be improved upon, or better yet, be designed ***de novo***."

Defining Terms

Ab initio The computational construction of a molecule from its constituent atoms without any prior assumptions other than the identity, quantum mechanical properties, and predetermined connectivity of those atoms.

Actinide A series of *f*-block elements with increasing numbers of nuclear protons (atomic numbers 89–104) from actinium to rutherfordium.

Activation energy The energy needed to pass over the transition state when going from reactants to products.

AMBER A force-field method name that is an acronym for *assisted model building with energy refinement*.

Anions Negatively charged species, in which the total number of electrons exceeds the total number of protons.

Anti The disposition of two objects in reference to a plane so that the two objects are on opposite sides of the plane.

Asymmetric Without symmetry; *asymmetric* and *chiral* are synonyms.

Axial The location of a group that is directly attached to a six-membered ring structure and lies in a plane roughly perpendicular to that of the six-membered ring.

B3LYP An acronym for *Becke 3 term, Lee, Yang, and Parr*, an advanced variant of the density functional method that includes gradient correction factors for electron correlation potential energy.

Basis sets A set or collection of functions used to describe the atomic orbitals that are combined in the generation of a molecular orbital description of a compound.

Born–Oppenheimer approximation The assumption that because electron motion far exceeds nuclear motion, the electron motion is essentially independent of nuclear motion, so that the wave function is separated into two parts: one for the nucleus and one for the electrons.

Chair structure A conformation of the global energy minimum structure for cyclohexane (C_6H_{12}) that resembles a lounge chair (with a headrest and footrest), in which all opposing sides of the structure are parallel.

Chime A program developed by MDL for visualization of structural models on the Web.

Chiral A synonym of *asymmetric*; an object that is not identical to its mirror image.

Conformations Different structural representations of the same compound that differ only in the twist and turn of bonds (without breaking the bonds).

Conjugation An interaction of electron density in two or more adjacent systems. In a valence bond approach, adjacent pi bonds (or nonbonded electrons with a pi bond), with the pi electrons located in parallel, coplanar orbitals that can overlap each other, allow for interaction of the pi systems. Such an interaction imparts greater thermodynamic stability to the system than would be manifest if the adjacent electronic systems did not overlap (interact with) each other.

Coordination number The number of ligands that surround and are bonded to the central metal atom in a complexed structure.

Covalent bonds These occur when the electrons are shared evenly (but not necessarily equally) by each of two bonded atoms.

***d*-block elements** The metallic elements in which highest-energy electrons are placed in the *d* orbitals of the atom.

De novo A *de novo* design is the construction of an entirely new structure based on new insights (such as computational models), rather than on the modification of an existing structure.

Density functional theory (DFT) Those *ab initio* methods that deal with the total electron density of the molecule.

Diastereomeric relationship A relationship between two objects that are not mirror images of each other but are still stereoisomers, meaning that they have different locations of atoms in space where the differences are not conformational. Diastereomers are expected to have different chemical and physical properties.

Diels–Alder reaction A reaction named after the two chemists who discovered it. It is a pericyclic process resulting in the formation of a cyclohexene structure from two species, one having a 2-pi electron system and the other having a conjugated 4-pi electron system.

Diffusion function A model for treatment of electrons that are at a considerable distance from the nucleus of the atom.

Effective core potential (ECP) A potential function that represents the nonvalence (nonbonding) electrons of an atom.

Electron configuration The manner in which electrons are added to atoms of increasing atomic number. The same protocol for electron configurations is applied to filling electrons into atomic and molecular orbitals.

Electron correlation The avoidance behavior that characterizes electrons; the electrons do not want to be near each other.

Electron spin resonance (ESR) A technique that uses the interaction between the nuclear magnetic spin of nuclei that exist as single electron (radical) species with an applied magnetic field to obtain spectroscopic data. ESR can be used to characterize the environment in which the atoms reside.

Electrostatic potential map The result of an interaction between a charge and the surface of a molecule, which reveals positions of affinity for both negative and positive species within the molecule.

Empirical methods Methods based on parameters assumed to be representative of atoms, bonds, and interactions in the target structure, employing classical mathematical relationships using those parameters to solve for structural energies.

Enantiomer The named relationship between two objects that have the same composition and connectivity but are mirror images of each other and are not identical. Enantiomers have different locations of atoms in space where the differences are not conformational. A classic example of an enantiomeric relationship is the set of right and left hands.

Equatorial Pertaining to the location of a group that is directly attached to a six-membered ring and lies in a plane roughly approximating that of the carbon atoms of that ring.

Excited state A state in which the ground state (low-energy) electronic configuration of the molecule has been altered so that one electron in the structure is elevated to a higher level of energy.

***f*-block elements** The lanthanide and actinide metallic elements in which highest energy electrons are placed into the f orbitals of the atom.

Formal oxidation When atoms react by either giving up or accepting electrons, the formal accounting of that process is denoted by the change in formal oxidation of an element relative to the number of electrons that the neutral atom would contain in a nonbonded state.

Frontier molecular orbitals The highest-energy molecular orbitals of a compound, typically taken as those that are integral to a conjugated system.

Group theory Mathematical methods commonly used to characterize the symmetry of compounds.

Half chair structure The global energy maximum structure for a cyclohexane (C_6H_{12}) compound, which resembles a chair with one end of the structure, either the headrest or the footrest, flattened out in a planar arrangement.

Hamiltonian The term H from the Schrödinger equation that serves as the operator for energy.

Hartree–Fock (HF) method A self-consistent *ab initio* procedure that models the total number of electrons in a system N by N wave functions. In Hartree–Fock, electron correlation is not considered, and the variation principle is assumed.

Heterogeneous Pertaining to a system of two or more different phases, such as a mixture of an insoluble solid and a liquid (e.g., sand in water) or a mixture of two insoluble liquids (e.g., oil and water).

Homogeneous Pertaining to a system of only one phase, such as a mixture of a soluble solid and a liquid (e.g., NaCl in water) or a mixture of two soluble liquids (e.g., ethanol and water).

Hückel calculations The mathematical treatment developed by Hückel to deal with compounds in which the electron interactions can be described as a matrix, with the sum of the squares of the coefficients of each column in the matrix representing electron density on one atom, and of each row the electron density on one orbital.

Hybridization A construct developed by Linus Pauling that became a central tool of valence bond theory to adopt blended atom orbitals (e.g., spherically symmetrical s orbitals and dumbbell-shaped p orbitals) to create hypothetical mixed atom orbitals (e.g., sp^3, which denotes a mixing of one part s and three parts p) in order to account for observed bond angles and lengths.

Isomers Compounds that have the same elemental composition but differ from each other in some other respect. The three major types of isomers are constitutional (atoms have different connectivity), conformational, and configurational (atoms occupy different locations in space, and those differences cannot be resolved without breaking and remaking bonds).

Kinetics The study of the rates of chemical reactions.

Lennard–Jones 12-6 potential This is used to describe weaker interactions, such as van der Waals forces, and has the functional form $V_{\mathrm{LJ}}(r) = 4\varepsilon[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}]$. The $\sim 1/r^{12}$ is a repulsive term that dominates at short distances, and the $\sim 1/r^{6}$ is an attractive term that dominates at large distances. The parameter ε is the depth of the well, and σ is related to the equilibrium bond distance $\sigma = 2^{1/6} r_e$.

Ligand An atom or a bonded grouping of atoms that forms a bond to a centrally located atom in a complex by donating its electrons to that atom.

Linear combination of atomic orbitals (LCAO) The process of combining the atomic basis functions to generate a wave function.

Many-body potential A potential energy function that includes many-body effects by changing the pair potential to incorporate interactions from nearby atoms.

Mechanism The description of bond-making and bond-breaking in a chemical process.

Merck molecular force field (MMFF) A version of a MM program developed by Merck Pharmaceutical Company to evaluate organic compounds of pharmacological potential.

Metallic Pertaining to elements that have a proclivity to give up valence electrons to other elements when combined with those elements to form compounds.

Metathesis Any reaction in which an atom (or group of atoms) in one reactant switches with another atom (or group of atoms) from another reactant in a chemical transformation.

Michaelis–Menton equation The expression of kinetics in biochemical systems is most often treated with the Michaelis–Menton equation in which a substrate (S) and an enzyme (E) react to form a product (P) by intermediate steps described as enzyme–substrate complexation (ES) and enzyme–product decomplexation, so that E + S are in equilibrium with ES, which is in equilibrium with EP, which is in equilibrium with E + P.

Molecular dynamics (MD) An application of a force-field method with classic equations of motion in a series of time steps. MD is useful to describe processes such as SIMS simulation or conformation searching.

Molecular mechanics (MM) A method that primarily uses a classical mechanics ball-and-spring approach to evaluate the structure of a compound.

Molecular orbitals Concept used to describe the distribution of electrons in a molecule when assuming that the orbitals no longer belong to the specific atoms of which the molecule is composed, but rather to the entire molecule.

Møller–Plesset method Electron correlation in *ab initio* methods that is based on perturbation theory.

Monte Carlo method A randomization of the three-dimensional locations of objects, generally used to seek a lower energy region that may not be accessible or evident by more regular manipulations of the same objects in a more orderly or logical manner.

Morse potential An empirical pair potential that describes the stretching of a chemical bond. The functional form is $V(r) = D_e \left(1 - e^{-\beta(r - r_e)}\right)^2$, where r_e is the equilibrium bond distance, D_e is the depth of the well, and β is related to the curvature near the minimum.

Nuclear magnetic resonance (NMR) A technique that uses the interaction of nuclear magnetic spin of certain nuclei (like ^{1}H and ^{13}C) with an applied magnetic field to obtain spectroscopic data that can be used to characterize the environment in which the atoms reside.

Nucleophilic behavior The preference for an electron-rich species to bond to the positive nucleus of another species, other than hydrogen.

Orbital A three-dimensional volume of space which is the probable location for finding electrons. If the orbital belongs to an atom, it is called an *atomic orbital*; if it belongs to the molecule, it is called a *molecular orbital*.

Pair potential A potential energy function that describes how the potential energy between a pair of atoms depends on the distance between them.

Pairwise additive assumption The assumption that the interaction between each pair of atoms is independent of the other atoms in the system. The total potential energy of the system is assumed to be equal to the sum of the interaction between each pair of atoms: $V_{tot}(r_1, r_2, \ldots, r_n) = \sum_i \sum_{j>i} V(r_{ij})$.

Pericyclic processes Processes (including the Diels–Alder reaction) with cyclic transition states that reflect the reorganization of sigma and pi bonds in the chemical transformation.

Pi electrons Those electrons positioned in orbitals that allow for side-to-side overlap. In valence bond theory, p orbitals are used most commonly to construct pi systems.

Potential energy function A mathematical function with a functional form that describes how the potential energy depends on the relative position of each atom in the system. The function represents the actual solution to the Schrödinger equation within the Born–Oppenheimer approximation for the physical system. The parameters of the functional form are fit to experimental data and also to data from *ab initio* calculations.

Quantum mechanics The physics of the very small. The rules of quantum mechanics govern atomic and molecular phenomena and determine the relative probabilities of electron locations.

R Denotes the spatial orientation of groups about a central atom in one of the pair of enantiomeric orientations, R and S.

Radical species Having a single, non-bonded electron.

Raman spectroscopy A technique for measuring a molecule's vibrational, rotational, or electronic energy, which depends largely on polarizability.

Relativistic effects As electrons get closer to the nucleus, their speeds approach the speed of light and the theory of relativity must be applied to them. This is especially important for atoms with larger atomic numbers.

S Denotes the spatial orientation of groups about a central atom in one of the pair of enantiomeric orientations, R and S.

Saturated A carbon compound is saturated if the carbons in the compound are bonded to as many hydrogen atoms as possible. Using the formula C_nH_{2n+2}, where n is the number of carbons, $2n + 2$ is the number of hydrogen atoms required for the compound to be classified as saturated.

Schrödinger equation Erwin Schrödinger (also spelled *Schroedinger*) developed the model of the hydrogen atom using wave functions upon which *ab initio* methods are founded. In this model, it is possible to solve for an electron's energy with great precision, but the electrons' locations can only be described probabilistically.

Secondary ion mass spectrometry (SIMS) Procedure that bombards a sample with primary ions, causing the sample to eject molecules and molecular fragments from the sample surface. The charged ejected particles (i.e., the secondary ions) are then analyzed by mass spectrometry.

Self-consistent field (SCF) A method of iterative refinement, starting from an arbitrary initial value and performing calculations of new values, replacing the initial value with the new values until the new (lower energy) and initial (higher energy) converge to an acceptable degree.

Semiempirical method A method that includes both quantum mechanics and empirical parameters to compute molecular structures and properties.

Sigma electrons Those electrons in a valence bond system that are positioned in end-to-end, overlapping atomic orbitals between two atoms. Typically, the atomic orbitals in such a bond have at least some percentage composition of s-type (symmetrically disposed around the atom) character.

Spin Electrons are described by many characteristic quantities, one of which is the spin quantum number, which can be either $+1/2$ or $-1/2$.

Stiff equations Sets of differential equations whose solutions vary over so great a span of values that their simultaneous integration is problematic by normal methods, such as Euler or Runge–Kutta.

Syn The disposition of two objects in reference to a plane so that the two objects are on the same side of the plane.

Thermodynamic Pertaining to the study or characteristics of energy flux in any chemical process.

Torsion The interaction of two bonds or groups about a central connecting bond (e.g., interactions of bonds A–B and C–D or groups A and D in a structure such as A-B-C-D).

Transition metals The collection of *d*- and *f*- block elements (including the actinides and lanthanides).

Transition states The high-energy structures (where the bonds are partially made or broken) that represent the lowest energy barrier that must exist between starting materials and products in reactions that create bonds.

Twist boat conformation A local energy minimum structure for cyclohexane (C_6H_{12}) that resembles a boat (with a bow, stern, and transom) whose sides are not parallel.

Valence bond theory A theory that describes molecules as a series of bonds made by the overlap and sharing of spin-paired electrons in the atomic orbitals between the atoms.

Van der Waals force The force between objects that arises from temporary electrostatic interactions of their surrounding electrons.

Variational principle Since an estimated wave function will never be equal to or lower than the actual energy, the result of a wave function calculation is used repeatedly as the approximation in successive calculations, until the approximate value converges with the calculated value.

Vibrational frequency The frequency of the vibration of atoms, which depends on the masses of the atoms that are bonded and the strength of the bond between them.

Vital force The theory that compounds from living systems, called *organic compounds*, can only be made by living systems, *in vivo* (in life), and not by humans, *in vitro* (in glassware).

Wave function The description of the probable distribution of electrons as a function of their x, y, z coordinates and spin.

ZINDO A program developed by Zerner that uses semiempirical quantum mechanical values to compute molecular spectra.

References

Allinger, N.L., Yuh, Y.H., Lii, J.-H. 1989. Molecular mechanics — the MM3 force-field for hydrocarbons. 1. *J. Am. Chem. Soc.*, 111: 8551–8566.

Amos, R.D. 1986. Calculation of polarizability derivatives using analytic gradient methods, *Chem. Phys. Lett.*, 124: 376–381.

Berry, J.I., Ewing, A.E., and Winograd, N. 2001. Biological Systems. In *ToF-SIMS: Surface Analysis by Mass Spectrometry*, Vickerman, J.C. and Briggs, D., Eds., SurfaceSpectraLtd and IMPublications, London, 595–626.

Borden, W.T., and Davidson, E.R. 1996. The importance of including dynamic electron correlation in *ab initio* calculations, *Acc. Chem. Res.*, 29: 67–75.

Brenner, D.W. 1990. Empirical potential for hydrocarbons for use in simulations of the chemical vapor deposition of diamond films, *Phys. Rev. B*, 42: 9458–9471.

Bur, S.K., Lynch, S.M., and Padwa, A. 2002. Influence of ground state conformations on the intramolecular Diels–Alder reaction, *Org. Lett.*, 4(4): 473–476.

Carloni, P., Rothlisberger, U., and Parrinello, M. 2002. The role and perspective of *ab initio* molecular dynamics in the study of biological systems, *Acc. Chem. Res.*, 35: 455–464.

Crutzen, P.J. 1996. My life with O3, NOx, and other YZOx compounds (Nobel Lecture), *Angew. Chem., Int. Ed. Engl.*, 35: 1758–1777.

Cundari, T.R. 2000. Computational studies of transition metal–main group multiple bonding, *Chem. Rev.*, 100: 807–818.

Fleming, I. 1976. *Frontier Molecular Orbitals and Organic Chemical Reaction*, John Wiley & Sons, New York.

Frenking, G., and Fröhlich, N. 2000. The nature of the bonding in transition-metal compounds, *Chem. Rev*. 100: 717–774.

Frieden, C. 1993. Numerical integration of rate equations by computer, *Trends Biochem. Sci.* 18: 58–60.

Garrison, B.J. 2001. Molecular Dynamics Simulations, the Theoretical Partner to State SIMS. In *ToF-SIMS: Surface Analysis by Mass Spectrometry,* Vickerman, J.C. and Briggs, D., Eds., SurfaceSpectraLtd and IMPublications, London, 233–257.

Garrison, B.J., Delcorte, A., and Krantzman, K.D. 2000. Molecule liftoff from surfaces, *Accts. Chem. Res.* 33: 69–77.

Gear, C.W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, NJ.

Gillespie, D.T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comp. Phys.*, 22: 403–434.

Gillespie, D.T. 1977. Exact stochastic simulation of coupled chemical reactions*, J. Phys. Chem.*, 81: 2340–2361.

Guirgis, G.A., Bell, S., Zheng, C., Durig, J.R. 2002. Infrared and Raman spectra, conformational stability, vibrational assignment, *ab initio* calculations and r0 structural parameters for N-methylpropargylamine, *Phys. Chem. Chem. Phys.*, 4: 1438–1450.

Guirgis, G.A., Zhu, X., Bell, S., Durig, J.R. 2001. Conformational analysis, barriers to internal rotation, *ab initio* calculations, and vibrational assignment of 4-fluoro-1-butyne, *J. Phys. Chem. A*, 105: 363–373.

Kibby, M.R. 1969. Stochastic method for the simulation of biochemical systems on a digital computer, *Nature*, 222: 298–299.

Li, J., and Bursten, B.E. 2001. The Electronic Structure of Organoactinide Complexes via Relativistic Density Functional Theory: Applications to the Actinocene Complexes $An(\eta^8\text{-}C_8H_8)_2$ (An = Th-Am). In *Computational Organometallic Chemistry*, Cundari, T. R., Ed., Marcel Dekker, New York.

Lipkowitz, K.B., D'Hue, C.A., Sakamoto, T., and Stack, J.N. 2002. Stereocartography: a computational mapping technique that can locate regions of maximum stereoinduction around chiral catalysts, *J. Am. Chem. Soc.*, 124: 14255–14267.

McAdams, H.H., and Arkin, A. 1997. Stochastic mechanisms in gene expression, *Proc. Natl. Acad. Sci.* 94: 814–819.

Mendes, P. 1993. GEPASI: a software package for modeling the dynamics, steady states and control of biochemical and other systems, *Comput. Applic. Biosci.*, 9: 563–571.

Mendes, P. 1997. Biochemistry by numbers: simulation of biochemical pathways with GEPASI 3, *Trends Biochem. Sci.,* 22: 361–363.

Mendes, P., and Kell, D.P. 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation, *Bioinformatics,* 14: 869–883.

Mohamed, T.A., Guirgis, G.A., Nashed, Y.E., Durig, J.R. 1999. Spectra and structure of silicon-containing compounds. XXV. Raman and infrared spectra, r0 structural parameters, vibrational assignment, and *ab initio* calculations of ethyl chlorosilane-Si-d2, *Struct. Chem.*, 10: 333–348.

Molina, M.J. 1996. Polar ozone depletion (Nobel Lecture), *Angew. Chem., Int. Ed. Engl.*, 35: 1778–1785.

Monard, G., and Merz, K.M. 1999. Combined quantum mechanical/molecular mechanical methodologies applied to biomolecular systems, *Acc. Chem. Res.*, 32: 904–911.

Nguyen, T.C., Ward, D.W., Townes, J.A., White, A.K., Krantzman, K.D., and Garrison, B.J. 2000. A theoretical investigation of the yield-to-damage enhancement with polyatomic projectiles in organic SIMS, *J. Phys. Chem. B*, 104: 8221–8228.

Polavarapu, P.L. 1990. *Ab initio* vibrational Raman and Raman optical activity spectra, *J. Phys. Chem.*, 94: 8106–8112.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. 1992. Integration of Ordinary Differential Equations. In *Numerical Recipes in Fortran, 2nd Ed.,* Cambridge University Press, Cambridge, Chapter 19.

Rowland, F.S. 1996. Stratospheric ozone depletion by chlorofluorocarbons (Nobel Lecture), *Angew. Chem., Int. Ed. Engl.*, 35: 1786–1798.

Siegbahn, P.E.M., and Blomberg, M.R.A. 2000. Transition-metal systems in biochemistry studied by high-accuracy quantum chemical methods, *Chem. Rev*. 100: 421–437.

Stuart, S., Tutein, A.B, and Harrison, J.A. 2000. A reactive potential for hydrocarbons with intermolecular interactions, *J. Chem. Phys*, 112: 6472–6468.

Todebush, P.M., Liang, G., and Bowen, J.P. 2002. Molecular mechanics (MM4) force field development for phosphine and its alkyl derivatives, *Chirality*, 14: 220–231.

Torrent, M., Solà, M., and Frenking, G. 2000. Theoretical studies of some transition-metal mediated reactions of industrial and synthetic importance, *Chem. Rev*. 100: 439–493.

Townes, J.A., White, A.K., Wiggins, E.N., Krantzman, K.D., Garrison, B.J., and Winograd, N. 1999. Mechanism for increased yield with the SF_5^+ projectile in organic SIMS: the substrate effect, *J. Phys. Chem. A*, 24: 4587–4589.

Van Stipdonk, M.J. 2001. Polyatomic Cluster Beams. In *ToF-SIMS: Surface Analysis by Mass Spectrometry,* Vickerman, J.C. and Briggs, D., Eds., SurfaceSpectraLtd and IMPublications, London, 309–346.

Venkatesan, H., Davis, M.C., Altas, Y., Snyder, J.P., and Liotta, D.C. 2001. Total synthesis of SR 121463 A, a highly potent and selective vasopressin V2 receptor antagonist, *J. Org. Chem.*, 66: 3653–3661.

Woods, R.J., Andrews, C.W., and Bowen, J.P. 1992. Molecular mechanical investigations of the properties of oxocarbenium ions. 1. Parameter development, *J. Am. Chem. Soc.,* 114: 850–858.

Zaric, R., Person, B., Krantzman, K.D., Garrison, B.J. 1998. Molecular dynamics simulations to explore the effect of projectile size on the ejection of organic targets from metal surfaces, *Int. J. Mass Spectrom. Ion Processes*, 174: 155–166.

Further Information

General Reviews

Cramer, C.J. 2002. *Essentials of Computational Chemistry: Theories and Models*, John Wiley & Sons, New York.

Leach, A.R. 2001. *Molecular Modelling: Principles and Applications, 2nd Ed.*, Pearson Education, Essex, England.

Lipkowitz, K.B., and Boyd, D.B., Eds. 1990–2002. *Reviews in Computational Chemistry*, Vols 1–18, John Wiley & Sons, New York.

Schleyer, P.R. (Editor-in-Chief) 1998. *Encyclopedia of Computational Chemistry*, 5 vols., John Wiley & Sons, New York.

Young, D. 2001. *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*, John Wiley & Sons, Inc, New York.

Web Sites

BerkeleyMadonna: http://www.berkeleymadonna.com
Chemical kinetics simulator (CKS): http://www.almaden.ibm.com/st/msim
Encyclopedia of Computational Chemistry: http://www.mrw.interscience.wiley.com/ecc/
Gepasi biochemical simulation: http://www.gepasi.org/
Journal of Chemical Education (*JCE*): http://JChemEd.chem.wisc.edu
KINSIM: http://www.wistl.edu/cflab/message.html
Kintecus: http://www.kintecus.com
NIH Center for Molecular Modeling: http://cmm.info.nih.gov/modeling/
Project SERAPHIM: http://ice.chem.wisc.edu/seraphim
Reviews in Computational Chemistry index: http://chem.iupui.edu/~boyd/rcc.html
STELLA: http://www.hps-inc.com/

33

Computational Astrophysics

Jon Hakkila
College of Charleston

Derek Buzasi
U.S. Air Force Academy

Robert J. Thacker
McMaster University

33.1 Introduction

Modern **astronomy/astrophysics** is a computationally driven discipline. During the 1980s, it was said that an astronomer would choose a computer over a telescope if given the choice of only one tool. However, just as it is impossible to separate "astronomy" from astrophysics, most astrophysicists would no longer be able to separate the computational components of astrophysics from the processes of data collection, data analysis, and theory. The links between astronomy and computational astrophysics are so close that a discussion of computational astrophysics is essentially a summary of the role of computation in all of astronomy. We have chosen to concentrate on a few specific areas of interest in computational astrophysics rather than attempt the monumental task of summarizing the entire discipline. We further limit the context of this chapter by discussing astronomy rather than related disciplines such as planetary science and the engineering-oriented aspects of space science.

33.2 Astronomical Databases

Physics and astronomy have been leaders among the sciences in the widespread use of and access to online access and data retrieval. This has most likely occurred because the relatively small number of astronomers is broadly distributed so that few astronomers typically reside in any one location; it also perhaps results because astronomy is less protective of data rights than other disciplines driven more by commercial

spin-offs. Astronomers regularly enjoy online access to journal articles, astronomical catalogs/databases, and software.

33.2.1 Electronic Information Dissemination

All major astrophysical journals are available electronically [Boyce et al. 2001]. Several electronic preprint servers are also available [Ginsparg 1996, Hanisch 1999]. Furthermore, the Astrophysical Data Service (ADS) [Kurtz et al. 2000] is the electronic index to all major astrophysical publications; it is accessible from a variety of mirror sites worldwide. The ADS is a data retrieval tool that allows users to index journal articles by title, keyword, author, astronomical object and even by text search within an article. The ADS is a citation index as well as a tool capable of accessing the complete text of articles and in many cases the original data, although it has not served its original purpose (as designed by NASA) of allowing for the integrated data management of all astrophysics missions.

Electronic dissemination plays an important role in the collection of astronomical data. Many variable sources (such as high-energy transients and peculiar **galaxies**) often exhibit changes on extremely short timescales. Follow-up observations require the community to rapidly disseminate source locations, brightnesses, and other pertinent information electronically (notification sometimes must go out to other observatories on timescales as short as seconds). There are a variety of electronic circulars available to the observational community. The primary source of these is the International Astronomical Union (http://www.iau.org/).

33.2.2 Data Collection

Astronomy is generally an observational rather than experimental science (although there are laboratory experiments that emulate specific astronomical conditions). Modern astronomical observations are made using computer-controlled telescopes, balloon instruments, and/or satellites. Many of these are controlled remotely, and some are robotic.

All telescopes need electronic guidance because they are placed on moving platforms; even the Earth is a moving platform when studying the heavens. From the perspective of the terrestrial observer, the sky appears to pivot overhead about the Earth's equatorial axis; an electronic drive is needed to rotate the telescope with the sky to keep the telescope trained on the object. However, the great weight of large telescopes is easier to mount altazimuthally (perpendicular to the horizon) than equatorially. A processor is needed to make real-time transformations from equatorial to altazimuth coordinates so that the telescope can track objects; it is also needed to accurately and quickly point the telescope. Precession of the equinoxes must also be taken into account; the direction of the Earth's rotational axis slowly changes position relative to the sky. Computers have thus been integrated into modern telescope control systems.

Additional problems are present for telescopes flown on balloons, aircraft, and satellites. The telescopes must either be pointed remotely or at least have the directions they were pointed while observing accessible after the fact. Flight software is generally written in machine code to run in real-time. Stability in telescope pointing is required because astronomical sources are distant; long observing times are needed to integrate the small amount of light received from sources other than the sun, moon, and planets.

As an example, we mention computer guidance on the Hubble Space Telescope (HST). HST has different sensor types that provide feedback to maintain a high degree of telescope pointing accuracy; the Fine Guidance Sensors have been key since their installation in 1997. The three Fine Guidance Sensors provide an error signal so that the telescope has the pointing stability needed to produce high-resolution images. The first sensor monitors telescope pitch and yaw, the second monitors roll, and the third serves both as a scientific instrument and as a backup. Each sensor contains lenses, mirrors, prisms, servos, beam splitters, and photomultiplier tubes. Software coordinates the pointing of the sensors onto entries in an extremely large star catalog. The guidance sensors lock onto a star and then deviations in its motion to a 0.0028-arcsecond accuracy. This provides HST with the ability to point at the target to within 0.007

arcseconds of deviation over extended periods of time. This level of stability and precision is comparable to being able to hold a laser beam constant on a penny 350 miles away.

To complicate matters, many telescopes are designed to observe in **electromagnetic spectral** regimes other than the visible. An accurate computer guidance system is needed to ensure that the telescope is correctly pointing even when a source cannot be identified optically, and that the detectors are integrated with the other electronic telescope systems. Many bright sources at nonvisible wavelengths are extremely faint in the visible spectral regime. Furthermore, telescopes must be programmed to avoid pointing at bright objects that can burn out photosensitive equipment, and sometimes must avoid collecting data when conditions arise that are dangerous to operation. For example, orbital satellites must avoid operating when over the South Atlantic Ocean — in a region known as the South Atlantic Anomaly, where the Earth's magnetic field is distorted — because high-energy ions and electrons can interact with the satellite's electronic equipment. This can cause faulty instrument readings and/or introduce additional noise.

There are other examples where computation is necessary to the data collection process. In particular, we mention the developing field of **adaptive optics**. This process requires the fast, real-time inversion of large matrices [Beckers 1993]. *Speckle imaging techniques* (e.g., [Torgerson and Tyler 2002]) also remove atmospheric distortion by simultaneously taking short-exposure images from multiple cameras.

Detectors and electronics must often be integrated by one computer system. Many interesting specific data collection problems exist that require modern computer-controlled instrumentation. For example, radio telescopes with large dishes cannot move, and sources pass through the telescope's field-of-view as the Earth rotates. Image reconstruction is necessary from the data stream because all sources in the dish's field of view are visible at any given time. Another interesting data collection problem occurs in infrared astronomy. The sky is itself an infrared emitter, and strong source signal-to-noise can only be obtained by constantly subtracting sky flux from that of the observation. For this reason, infrared telescopes are equipped with an oscillating secondary mirror that wobbles back and forth, and flux measurements alternate between object and sky. A third example concerns the difficulty in observing x-ray and gamma-ray sources. X-ray and gamma-ray sources emit few photons, and these cannot be easily focused due to their short wavelengths. Computational techniques such as Monte Carlo analysis are needed to deconvolve photon energy, flux, and source direction from the instrumental response. Very often, this analysis must be done in real-time, requiring a fast processor. In addition, the telescope guidance system must be coordinated with onboard visual observations because the visual sky still provides the basis for telescope pointing and satellite navigation.

33.2.3 Accessing Astronomical Databases

Database management techniques have allowed astronomers to address the increasing problem of storing and accurately cross-referencing astronomical observations. Historically, bright stars were given catalog labels based on their intensities and on the constellation in which they were found. Subsequent catalogs were sensitive to an increased number of fainter objects and thus became larger while simultaneously assigning additional catalog numbers or identifiers to the same objects. The advent of photographic and photoelectric measurement techniques and the development of larger telescopes dramatically increased catalog sizes. Additional labels were given to stars in specialty catalogs (e.g., those containing bright stars, variable stars, and binary stars). Solar system objects such as asteroids and comets are not stationary and have been placed in catalogs of their own.

In 1781, Charles Messier developed a catalog of fuzzy objects that were often confused with comets. Subsequent observations led to identification of these extended objects as star clusters, galaxies, and gaseous nebulae. Many of these extended astronomical sources (particulary regions of the **interstellar medium**) do not have easily identified boundaries (and the observed boundaries are often functions of the **passband** used); this inherent fuzziness presents a problem in finding unique identifiers as well as a database management problem. Furthermore, sources are often extended in the radial direction (sources are three-dimensional), which presents additional problems because distances are among the most difficult astrophysical quantities to measure.

As astronomy entered the realm of multi-wavelength observations in the 1940s, observers realized the difficulty in directly associating objects observed in different spectral regimes. An x-ray emitter might be undetectable or appear associated with an otherwise unexciting stellar source when observed in the optical. Angular resolution is a function of wavelength, so it is not always easy to directly associate objects observed in different passbands.

Temporal variability further complicates source identification. Some objects detected during one epoch are absent in observations made during other epochs. Signal-to-noise ratios of detectors used in each epoch contribute to the source identification problem. Additionally, gamma-ray and x-ray sources tend to be more intrinsically variable due to the violent, nonthermal nature of their emission. Examples of sources requiring access via their temporal characteristics include supernovae, gamma-ray bursts, high-energy transient sources, and some variable stars and extra-galactic objects.

There are tremendous numbers of catalogs available to astronomers, and many of these are found online. Perhaps the largest single repository of online catalogs and metadata links is VisieR (http://vizier.u-strasbg.fr/viz-bin/VizieR) [Ochsenbein et al. 2000]. Online catalogs also exist at many other sites, including HEASARC (High Energy Astrophysics Science Archive Research Center at http://heasarc.gsfc.nasa.gov/), HST (Hubble Space Telescope at http://www.stsci.edu/resources/), and NED (NASA/IPAC Extragalactic Database at http://nedwww.ipac.caltech.edu/).

Large astronomical databases exist for specific ground-based telescopes and orbital satellites. Some of these databases are large enough to present information retrieval problems. Examples of these databases are 2MASS (Two Micron All Sky Survey) [Kleinmann et al. 1994]; DPOSS (Digitized Palomar Observatory Sky Survey) [Djorgovski et al. 2002]; SDSS (Sloan Digital Sky Survey) [York et al. 2000]; and NVSS (The NRAO VLA Sky Survey) [Condon et al. 1998]. Databases span the range of astronomical objects from stars to galaxies, from active galactic nuclei to the interstellar medium, and from gamma-ray bursts to the cosmic microwave background. Databases are often specific to observations made in predefined spectral regimes rather than specific to particular types of sources; this reflects the characteristics of the instrument making the observations.

33.2.4 Data File Formats

The astronomic community has evolved a standard data format for the transfer of data. The Flexible Image Transport System (FITS) has broad general acceptance within the astronomic community and can be used for transferring images, spectroscopic information, time series, etc. [Hanisch et al. 2001]. It consists of an ASCII text header with information describing the data structure that follows. Although originally defined for nine-track, half-inch magnetic tape, FITS format has evolved to be generic to different storage media. The general FITS structure has undergone incremental improvements, with acceptance determined by vote of the International Astronomical Union.

Despite the general acceptance of FITS file format, other methods are also used for astronomical data transfer. This is not surprising, given the large range of data types and uses. Some data types have been difficult to characterize in FITS formats (such as solar magnetospheric data). Satellite and balloon data formats are often specific to each instrument and/or satellite.

Due to the large need for storing astronomical images, a wide range of image compression techniques have been applied to astronomy. These include fractal, wavelets, pyramidal median, and JPEG (e.g. [Louys et al. 1999]).

33.3 Data Analysis

Mathematical and statistical analyses are the driving forces behind the use of computation in astrophysics.

Data analysis and theoretical software can be accessed at a variety of sites worldwide. A few of the most well-known sites include the Astrophysics Source Code Library (http://ascl.net/), the UK Starlink site (http://star-www.rl.ac.uk/), and the Astronomical Software and Documentation Service at STScI, (http://asds.stsci.edu). Data analysis tools written in RSI's proprietary IDL programming language are

available at the IDL Astronomy User's Library (http://idlastro.gsfc.nasa.gov/homepage.html). IDL has become a computing language of choice by many astronomers because it has been designed as an image-processing language, it has been written with mathematical and statistical uses in mind, it can handle multidimensional arrays easily, and it has many built-in data visualization tools.

33.3.1 Data Analysis Systems

Some 20 years ago, a myriad of different data analysis systems existed within astronomy. Typically, each institution (or sometimes individual groups within an institution) had its own data analysis package, and compatibility between these packages was limited or nonexistent. In the fall of 1981, however, astronomers at Kitt Peak National Observatory began development of the Image Reduction and Analysis Facility (IRAF), intended to serve as a general-purpose, flexible, extendable data reduction package. IRAF has grown beyond its original use primarily by ground-based optical astronomers to encompass space-based experiments as well at wavelengths ranging from x-ray to infrared.

IRAF is currently a mature system, with new releases occurring roughly annually, and is operated by about 5000 users at 1500 distinct sites. It is supported under a number of different computer architectures running UNIX or UNIX-like (e.g., Linux) operating systems. Oversight of IRAF development is formalized, with a Technical Working Group and various User's Committees overseeing evolution of the software. Numerous large astronomical projects have adopted IRAF as their data analysis suite of choice, typically by providing extensions to the basic system. These projects include the x-ray "Great Observatory" Chandra, the Hubble Space Telescope, PROS (ROSAT XRAY Data Analysis System), and FTOOLS (a FITS utility package from HEASARC).

The IRAF core system provides data I/O tools, interactive graphics and image display tools, and a variety of image manipulation and statistical tools. Commonly available "packages" that are part of the standard installation include tools for basic CCD data reduction and photometry, and support one-dimensional, two-dimensional, echelle, and fiber spectroscopy. Most tasks can be operated in either interactive or batch mode.

IRAF supports the FITS file format, as well as its own internal data type. In addition, extensions exist to handle other data types. Currently, those include STF format (used for HST data), QPOE format (used for event list data such as from the x-ray and EUV satellites), and PLIO for pixel lists (used to flag individual pixels in a region of interest). IRAF uses text files as database or configuration files, and provides a number of conversion tools to produce images from text-based data. Binary tables can be manipulated directly using the TABLES package.

Other general-purpose data reduction packages coexist with IRAF. These include the Astronomical Image Processing System (AIPS), and its successor (AIPS++), developed at the U.S. National Radio Astronomy Observatory (NRAO), and is still the image analysis suite of choice for radio astronomy. Image processing has been a very important subdiscipline within computational astrophysics, and we mention a number of image reconstruction methods with special applications to astronomy: the Maximum Entropy Method (e.g., [Lasenby et al. 2001]; the Pixon Method [Piña and Puetter 1993]; and Massive Inference [Skilling 1998]). XIMAGE and its relatives (XSPEC and XRONOS) serve a similar function for the x-ray astronomy community. A significant number of optical astronomers use Figaro, developed at the Anglo-Australian telescope. Figaro is particularly popular throughout Australia and the United Kingdom. Recently, Figaro has been adapted to run within the IRAF system, allowing users to have the best of both worlds. While there is no formal software system operating under IDL, the profusion of astronomical programs available in that proprietary language makes it worthy of mention.

One of the most significant drivers for development of mission-independent data analysis software within astronomy has been NASA's Applied Information Systems Research Program (AISRP). This program has encouraged the development of software to serve community-wide needs and has also fought the recurrent tendency for each project to develop its own software system, but instead to write software "packages" within the IRAF or IDL architecture. Of course, a general data analysis system is not the best solution for all specialized needs, particularly for space-based astronomy. In these cases, many missions have developed their own packages, and not all of these exist within an IRAF/AIPS/XIMAGE/IDL framework.

33.3.2 Data Mining

For many years, both NASA and ESA (the European Space Agency) have collected and preserved data from observatories in space. Similar activities are underway (although with varying degrees of success) at ground-based observatories. Most of these archives are available online, although the heterogenous nature of user interfaces and metadata formats — even when constrained by HTML and CGI — can make combining data from multiple sources an unnecessarily involved process. In addition, astronomical archives are growing at rates unheard of in the past: terabytes (TB) per year are now typical for most significant archives. Two recently established databases that illustrate the trend are the MACHO database (8 TB) and the Sloan Digital Sky Survey (15 TB).

The simplest kinds of questions one might ask of these kinds of data sets involve resource discovery; for example, "Find all the sources within a given circle on the sky," or "List all the observations of a given object at x-ray wavelengths." Facilities such as SIMBAD [Wenger et al. 2000]; VisieR [Ochsenbein et al. 2000]; NED [Mazzarella et al. 2001]; and ASTROBROWSE [Heikkila et al. 1999] permit these kinds of queries, although they are still far from comprehensive in their selection of catalogs and/or data sets to be queried. One problem arises from the nature of the FITS format, which is poorly defined as far as metadata are concerned; XML may provide a solution here, but astronomers have yet to agree on a consistent, common set of metadata tags and attributes.

Difficulties due to heterogeneous formats and user interfaces are being addressed by a number of so-called **virtual observatory** projects, such as the U.S. National Virtual Observatory (NVO), the European Astrophysical Virtual Observatory (AVO), and the British Astrogrid Consortium. The most basic intent of all these projects, which are coordinated with one another at some level, is to deliver a form of integrated access to the vast and disparate collection of existing astronomical data. In addition, all intend to provide data visualization and mining tools.

In its most encompassing form, astronomical data mining involves combining data from disparate data sets involving multiple sensors, multiple spectral regimes, and multiple spatial and spectral resolutions. Sources within the data are frequently time-variable. In addition, the data are contaminated with ill-defined noise and systematic effects caused by varying data sampling rates and gaps and instrumental characteristics. Finally, astronomers typically wish to compare observations with the results of simulations, which may be performed with mesh scales dissimilar to that of the observations and which suffer from systematic effects of their own. It has been observed [Page 2001] that data mining in astronomy presently (and in the near future) focuses on the following functional areas:

1. Cross-correlation to find association rules
2. Finding outliers from distributions
3. Sequence analysis
4. Similarity searches
5. Clustering and classification

Despite the difficulties outlined above, the nature of astronomical data — which are generally freely available and in computer-accessible form — has led to numerous early applications of data mining techniques in the field. As far back as 1981, FOCAS (Faint Object Classification and Analysis System) [Jarvis and Tyson 1981] was developed for the detection and classification of images on astronomical plates for the automatic assembly of catalogs of faint objects. Neural nets and decision trees have also been applied for the purposes of discriminating between galaxies and stars in image data [Odewahn 1992, Fayyad 1996] and for morphological classification of galaxies [Storrie-Lombardi 1992].

More recently, projects such as JPL's Diamond Eye [Roden et al. 1999] have begun to experiment with more general data mining enterprises. In this particular case, users interact with data mining servers via a Java applet (and thus need not have any particular data mining expertise themselves). Algorithms being tested include adaptive recognition (and its application to dynamic events) and ad hoc queries. Another current program is SKICAT (http://www-aig.jpl.nasa.gov/public/mls/skicat/skicat_home.html), which is an integrated software system applying image processing, database management, and AI classification

to large image database analysis. The basic image processing routines detect objects and measure a set of features (surface brightness, extent, morphology, etc.) for each object. Algorithms such as GID3*, O-BTree, and RULER are used to produce decision trees and classification rules based on training data, and these classifiers are then applied to the new data.

The greatest near-term successes of data mining are likely to arise from its application to large but coherent data sets. In this case, the data has a common format and noise characteristics, and data mining applications can be planned from the beginning. Perhaps the most ambitious of such ongoing projects are the Sloan Digital Sky Survey (SDSS) [York et al. 2000] and 2MASS (e.g. [Nikolaev et al. 2000]). SDSS uses a dedicated 2.5-meter telescope to gather images of approximately 25% of the sky (some 10^8 objects), together with spectra of approximately 10^5 objects of cosmological significance. Pipelines have been developed to convert the raw images into astrometric, spectroscopic, and photometric data that will be stored in a common science database, which are indexed in a hierarchical manner. The science database is accessible via a specialized query engine. The SDSS has required Microsoft to put new features into its SQL server; in particular, Microsoft has added a tree structure to allow rapid processing of queries with geographic parameters. The Two Micron All Sky Survey (2MASS) is another high-resolution infrared sky survey that contains massive amounts of data for discrete as well as nebulous sources.

The key problem in astronomical data mining will most likely revolve around interpretation of discovered classes. Astronomy is ruled by instrumental and sampling biases; these systematic effects can cause data to cluster in ways that mimic true source populations statistically. Because data mining tools are capable of finding classes that are only weakly defined statistically, these tools enhance the user's ability to find "false" classes. Subtle systematic biases are present (but minimized) even in controlled cases when data are collected from only one instrument. The astronomical community must be careful to test the hypothesis so that class structures are not produced by instrumental and/or sampling biases before accepting the validity of newly discovered classes.

33.3.3 Multi-Wavelength Studies

Multi-wavelength studies have become increasingly important in astronomy, as new spectral regimes (x-ray, extreme UV, and gamma ray) have opened to practitioners, and astronomers have become aware that most problems cannot be adequately addressed by studies within any one spectral band. Such studies are now recognized as essential to the understanding of objects ranging from the Sun and stars, through x-ray bursters and classical novae, to the interstellar medium and active galactic nuclei. The computational requirements peculiar to multi-wavelength astrophysics essentially fall into one of two categories, where the distinction is in the time rather than the spectral domain.

In the first case, we have situations in which the timescales associated with the phenomena under study are long compared with typical observational timescales. One such case is in studies of the interstellar medium (ISM), which is important in astrophysics because galactic gas and dust clouds are the source of new generations of stars. The ISM becomes more enriched as generations of stars die and return heavy elements to it. In the study of the ISM, one can identify three characteristic temperature scales: <100 K, $\approx 10^4$ K, and $\approx 10^7$ K; these temperatures necessitate observations at radio/infrared, optical/UV, and x-ray wavelengths, respectively (e.g., [Zhang et al. 2001]). In each case, the strength of its emission or absorption depends on the local density, composition, metallicity, temperature, and distribution of ambient photons. Because this causes the interstellar medium (and to a lesser extent the intergalactic medium) to interfere with observations of other galactic and extragalactic sources, stellar, galactic, and extragalactic astronomers are interested in knowing the radiative properties of the interstellar medium, which is by definition nebulous and three-dimensional, as well as knowing where this material is located. An example of code used to locate the interstellar medium in the visual spectral regime can be found at http://ascl.net/extinct.html [Hakkila et al. 1997]. Because the characteric evolutionary timescale of the ISM is long by human standards, simultaneous (and even contemporaneous) multi-wavelength observations are unnecessary. In this milieu, catalogs and databases come into their own, and archival multi-wavelength research is possible, focusing on spatial rather than temporal correlations.

An idea of the numerous databases and collections of online data available on the ISM can be found at http://adc.gsfc.nasa.gov/adc/quick_ref/ref_ism.html.

A different situation occurs with variable sources such as gamma-ray bursters, and extremely short-duration and high-energy events occurring at cosmological distances. The source of the bursts is as yet unknown, and they may be created by mergers of a pair of neutron stars or black holes, or by a hypernova, a type of exceptionally violent exploding star. Gamma-ray bursts and their afterglows have been detected across the electromagnetic spectrum, and further study of these objects clearly calls for multi-wavelength studies (e.g. [Galama 1999]). However, unlike the case obtaining for the ISM, gamma-ray bursts have timescales ranging from milliseconds up to approximately 10^3 seconds, and thus coordinated and simultaneous (or near-simultaneous) multi-wavelength observations are essential. In this case, the computational demand is more on rapid deployment of various computer-controlled telescopes (both on the ground and in space) than on correlation analyses of existing databases. Thus, systems such as GCN (GRB Coordinates Network) have been developed [Barthelmy et al. 2001] to distribute locations of GRBs to observers in real-time or near-real-time, and to distribute reports of the resulting follow-up observations.

33.3.4 Specific Examples

33.3.4.1 Cosmic Microwave Background (CMB) Data Analysis

One of the main goals of CMB data analysis is to derive the power spectrum of the temperature fluctuations, which correlates directly to fluctuations in the density of matter in the early Universe. The precise spectrum of matter perturbations depends acutely on cosmological parameters and, hence, CMB data is an excellent diagnostic for determining the parameters of cosmological models. However, the task of constructing the power spectrum is daunting because the raw CMB signal has instrument noise, as well as noise from the interstellar medium and other astrophysical objects, imposed upon it. To further complicate matters, all these sources of noise are often correlated.

Analysis proceeds by first creating a physical map from the time series of instrument pointings. A pixel map is first constructed by dividing up the area of sky surveyed. For the COBE experiment [Smoot et al. 1992], only a few thousand pixels were necessary; while for the PLANCK Surveyor satellite (http://astro.estec.esa.nl/SA-general/Projects/Planck/), tens of millions of pixels are required. Thus, the main step in creating the map is the separation of noise, which can be done under the assumption that the time-series of noise signals is drawn from a Gaussian distribution. Linear algebra methods are used to calculate the pixel-pixel noise correlation, which is then used to construct the map. Brute-force methods are inefficient and exploiting sparse matrix methods is the only way to do the calculation efficiently. Nonetheless, the calculation is sufficiently large that massively parallel computing is necessary, and a general analysis package (MADCAP) has been developed [Borrill 1999].

Having constructed the map, the next step is to calculate the power spectrum. This is a significantly more difficult process than map creation, because each pixel contains information about the signal on all scales within the map. Further, the power spectrum that produces the signal must be derived from a maximum likelyhood analysis, which in turn requires an iterative process. The steps involved in the iterative process are extremely computationally costly because correlation matrices and numerical derivatives must be calculated multiple times. Estimates of the total number of flops required to analyze the PLANCK data set are close to 10^{24}, which would take almost 1000 years on the Earth Simulator alone. Even smaller data sets, such as MAXIMA-2, require tens of petaflops. Ultimately, analysis of the larger data sets requires the development of new analysis methods.

33.3.4.2 Gamma-Ray Burst Data Analysis

Gamma-ray bursts (GRBs) are short bursts of primarily gamma-radiation having fluxes and spectra that evolve on short timescales. Evidence suggests that GRBs are produced when relativistic shock waves collide with each other and with the ambient interstellar medium. The source of the shocks is currently presumed to be a *hypernova*; a supernova variant in which a significant portion of the collapsing stellar core's energy is focused into shocked bipolar accelerated particle beams. GRBs exhibit a wide variety of complex behaviors

and yet there is evidence of multiple GRB classes. Class identification is as important in astrophysics as in other sciences; the properties of distinct GRB classes can lead to a better general understanding of GRB physics as well as to a better understanding of the different sources and environments producing the classes. However, class identification is not helpful if the mechanisms responsible for the producing the classes cannot be determined.

Data mining techniques have proven useful in the study of GRB classes. Data mining techniques are needed to identify clusters in the GRB attribute space because individual GRB behaviors overlap. Some overlap results from the large intrinsic range of GRB behaviors, some is due to distinctly different properties of GRB classes, and some is due to observational and instrumental bias. Bias can cause phantom classes to appear by creating clustering where no distinct source populations exist.

A reference to data mining techniques applied to GRBs can be found in [Hakkila et al. 2003]. GRB data mining has thus far been used primarily with the large GRB database collected by BATSE (the Burst And Transient Source Experiment on NASA's Compton Gamma-Ray Observatory) because this experiment has well-documented properties [Paciesas et al. 1999]. Data mining tools identify three distinct GRB classes rather than two known historically. Detailed analyses find that the newest (third) GRB class does not represent a separate source population; it is, instead, produced by observational biases resulting primarily from low signal-to-noise observations and from the instrumental trigger characteristics [Hakkila et al. 2003]. This successful result indicates that scientific data mining can be used not just to determine that classes exist, but also to determine *why* they exist.

33.3.4.3 Time Series Analysis

Astronomers make use of time series analysis techniques for a variety of purposes, including studies of variable stars and cataclysmic variables [Gilliland et al. 1998, Kiss et al. 2001]; pulsating or oscillating stars [Buzasi et al. 2000, Poretti et al. 2002]; asteroid rotation rates, active galactic nuclei [Pronik et al. 1999]; and detection of extrasolar planets [Brown et al. 2001]. Typically, astronomical time series suffer from relatively low signal-to-noise ratios, uneven sampling, and numerous gaps, at times leading to severe aliasing at the 1 day^{-1} frequency and difficulties in estimating a traditional autocorrelation function. In addition, some applications (e.g., AGN studies and planetary detection) are dominated by nonsinusoidal signals or pulses.

Historically, the primary tools used to support these efforts have been discrete Fourier transforms and periodograms [Scargle 1982]. Both of these estimators of the autocorrelation function can be defined in such a way as to satisfactorily represent unevenly sampled data, and small gaps in the time series can be handled using clever binning or interpolation techniques, but these necessarily distort the data and lead to the loss of information. Early efforts to deal with the difficulties raised by irregular sampling and gaps focused on applying the CLEAN algorithm [Roberts et al. 1987] to apply a nonlinear deconvolution in the frequency domain. Unfortunately, CLEAN, originally developed for use by radio astronomers, suffers from nonuniqueness as well as the tendency to fail when aliasing problems are severe. In some cases, aliasing can be minimized or eliminated by experimental design (e.g., GONG), but more often astronomers are confronted with this fundamental problem and this has driven numerous recent efforts in the area.

Recent developments in this area focus on the application of nonlinear, nonstationary techniques and Bayesian methods. The wavelet transformation shows great promise [Abry et al. 1995; Scargle 1997] because, unlike the DFT, it can be used to construct power spectrum estimators that are nearly independent of the signal shape and amplitude in the presence of noise. Such estimators are likely to find increasing use in the planetary-detection and AGN modeling communities. Perhaps an even more useful application of wavelets is in the denoising of power spectra, a technique pioneered by the solar physics community. Bayesian methods are also increasing in popularity, as they are well-suited to finding change points in long series of time-tagged data such as is typically obtained from high-energy astrophysics experiments [Scargle 1998].

A rapidly growing difficulty is the size of power spectra produced by astronomical experiments. Helioseismological observations can easily give rise to time series with in excess of 10^7 points, and asteroseismological observations are rapidly approaching this level [Schou and Buzasi 2001]. Achieving the maximum time resolution inherent in data such as time-tagged x-ray photon lists can require estimators with in excess

of 10^9 points [Ransom et al. 2002]. Furthermore, upcoming space-based experiments such as Eddington and Kepler will give rise to data sets that are so large as to mandate the use of automated techniques.

33.4 Theoretical Modeling

Prior to the advent of computers, theoretical modeling consisted largely of solving idealized systems of equations for a given problem. Very often, to make problems tractable, simplifying assumptions such as spherical symmetry and linearization of the problem would be necessary. Rapid numerical solutions of equations avoid the need for simplifying assumptions, but at a cost: one can no longer achieve an elegant formula for the solution to the problem, and insight often derived from manually solving the equations is lost.

33.4.1 The Role of Simulation

Although computation is commonplace in theoretical modeling, perhaps the most heavily computationally biased aspect is *simulation*. Simulation can be viewed as an extension of finding numerical solutions to a given equation set; however, the set of equations is often enormous in size (such as that produced by the gravitational interaction problem between a large number of bodies). Simulation also often involves visualizing the "data set" to help understand the phenomenon being studied, and the systems under investigation are almost always cast as an Initial Value Problem.

The roots of simulation in astrophysics can be traced back to at least the 1940s. Driven by a desire to understand the clustering of galaxies, Holmberg built an analog computer consisting of light sources and photocells to simulate the mutual interaction of two galaxies via gravity. Today, almost all simulation is conducted on digital computers. Development of fast, efficient algorithms for solving complex equation sets can often lead to programs containing tens of thousands of lines of code. Although it has been the tradition for individual researchers to develop codes in isolation, the past few years have seen the appearance of collaborations of researchers who work together on large coding projects. This trend is likely to continue and it is probable that in the near future researchers will converge to using a handful of readily available simulation packages (e.g., NEMO, http://bima.astro.umd.edu/nemo/).

33.4.2 The Gravitational *n*-body Problem

Although Newton's Law of Universal Gravitation has been supplanted by **General Relativity**, Newton's Law remains highly accurate for a very large number of astrophysical problems. However, solving the interaction problem for any number of bodies (n bodies) is difficult because at first appearances, the number of operations scales as n^2. However, provided that small errors in the force calculation are acceptable (RMS errors typically less than 0.5%), then approximate solutions can be found using order $n \log n$ operations. Roughly speaking, the algorithms used by researchers fall into two categories: treecodes [Barnes and Hut 1986] and grid (FFT) methods [Hockney and Eastwood 1981]. Treecodes are usually about an order of magnitude slower than grid codes for homogeneous distributions of particles, but are potentially much faster for very inhomogeneous distributions. To date, the largest gravitational simulations conducted contain approximately 1 billion particles, and have been used to coarsely simulate volumes representing as much as 10% of the entire visible universe [Evrard et al. 2002].

33.4.3 Hydrodynamics

Hydrodynamic modeling — or equivalently, computational fluid dynamics — plays an extremely important role in astrophysics. Although most astrophysical **plasmas** are not fluids in the everyday sense, the physical description of them is the same. Modern hydrodynamic methods fall into two main groups: Eulerian (fixed) descriptions and Lagrangian (moving) descriptions. Eulerian descriptions can be broadly decomposed into finite difference and finite element methods. In astrophysics, the finite difference method is by far the most common approach. Lagrangian descriptions can be decomposed into (1) "moving mesh"

methods where the grid deforms with the flow in the fluid, and (2) particle methods, for which Smoothed Particle Hydrodynamics (SPH) is a popular example [Gingold and Monaghan 1977].

Because shocks play an important role in the evolution of stars and the ISM, a significant amount of research has focused on "shock capturing methods." Most early approaches to shock capturing, and indeed a number of methods still in use today, provide stability by using an artificial viscosity to smooth out flow discontinuities (shocks). Although these methods work well, they often introduce additional, unwanted dissipation into the simulation. Perhaps the best alternative approach is the Godunov's Method [Godunov 1959], which is a simple example of a first-order method where the Riemann shock tube problem is solved at the interface of each grid cell. More modern algorithms have extended this idea to higher-order integration schemes, such as the Piecewise-Parabolic Method [Collela and Woodward 1984].

33.4.4 Magnetohydrodynamics and Radiative Transfer

Magnetic fluid dynamic modeling (MHD) is the focus of a large amount of research in computational astrophysics [Falgarone and Passot 2003]. The system of equations for MHD is that of hydrodynamics plus the addition of coupling terms corresponding to magnetic and electric forces and Maxwell's equations that constrain and evolve the magnetic and electric fields. Because of the severe complexities arising from the divergenceless nature of the magnetic field, most MHD methods are finite difference; and although particle methods have been used, quite often they produce significant integration errors.

Modern methods, as in hydrodynamics, often cast the problem in terms of a system of conservation laws. It is usual to look at variation along a given axis direction and to recast the problem in terms of "characteristic variables" that are constructed from eigenvalues and the primitive variables, such as density, pressure, and flow speed. Such recasting aids the development of the numerical method because timestepping can be viewed as propagating the system an infinitesimal amount along a characteristic. This formulation often allows development of stable integration schemes that produce accurate numerical solutions even when large time steps are used.

Radiative transfer (RT), the study of how radiation interacts with gaseous plasmas, is an extremely difficult problem. It bears parallels to gravity in that all points within a system can usually affect all others, but is further complicated by the possibility of objects along any given direction producing non-isotropic attenuation. The radiation intensity is a function of position, two angles for direction of propagation, and frequency — a total of six independent variables. There any many different approaches to solving RT, ranging from explicit ray tracing to Monte Carlo methods, as well as characteristic methods [Peraiah 2001]. Much of the modern research effort focuses on deriving useful approximation methods that ease the computational effort required.

33.4.5 Planetary and Solar System Dynamics

Recent advances in telescope instrumentation have led to a cascade of discoveries of extra-solar planets and, at the time of writing, more than 100 extra-solar planets are known. Consequently, there is now a large amount of interest in studying planet and solar-system formation. Solar-system formation occurs during star formation, and the inherent differences in the planets are due to a differentiation process that enables different elements to condense out of the solar nebula at different radii.

Planets form by hierarchical merging processes within the disk of the solar nebula. Dust grains form the first level of the hierarchy and planets the last, while objects of all mass scales and sizes exist in between. It should be noted that representing this variation of masses and sizes within a simulation is impossible because resolution is always limited by the available computing power and memory. Theoretical models of the agglomeration process must include hydrodynamics and gravity, although currently there is debate about the role of hydrodynamics in gas giant planet formation. At present, theory can be roughly divided into two approaches: (1) the study of stability properties of the solar nebula disk from an analytic perspective, and (2) the simulation of the process from a first principles perspective. Simulations with a million mass elements, designed to follow the agglomeration process in the inner part of the solar system,

were conducted in 1998 [Richardson et al. 2000]. More recent hydrodynamic simulations [Mayer et al. 2002] using the SPH technique have shown that gas giant planets can form extremely rapidly because of instabilities in protoplanetary disks.

The realization that our Solar System contains many small asteroids and meteorites capable of causing severe damage to the Earth has renewed interest in solar system dynamics. Calculating accurate orbits for these systems is difficult because they often have chaotic orbits. Chaotic systems place great demands on numerical integrations because truncation errors can rapidly pollute the integration. Thus, the integration schemes used must be highly accurate (often quadruple precision is used), and much effort has been devoted to "long-term" integrators (such as "sympletic integrators," see [Clarke and West 1997]) that preserve numerical stability over long-periods of simulation time. The chaos observed in long-term simulations of the solar system inspired a new theory [Murray and Holman 1999] that demonstrates that the Solar System is chaotic (Uranus could possibly be ejected) but the timescale for this is extremely long (10^{17} years).

33.4.6 Stellar Astrophysics

Among the first astrophysical problems to be addressed using modern computational methods were models of stellar interiors [Henyey et al. 1959; Cox et al. 1960] and atmospheres [Kurucz 1969]. One simplifying assumption needed for early stellar codes was that of Local Thermodynamic Equilibrium, which meant that stellar structural variations were not expected to occur on short timescales. Such assumptions are no longer necessary: stellar interior and atmosphere codes have become increasingly complex as computers and computational techniques have evolved. Theoreticians have been able to study rapid evolutionary phases and complex atmospheric processes in stars. Some of the difficult problems currently being addressed include the evolution of rotating stars [Meynet and Maeder 2002]; radial and nonradial stellar pulsations [Buchler et al. 1997; Crsico and Benvenuto 2002]; stellar magnetospheres [Wade et al. 2001]; evolution of stars in binary systems [Beer and Podsiadlowski 2002]; and supernovae [Woosley et al. 2002].

33.4.7 Star Formation and the Interstellar Medium

One of the greatest challenges in astrophysics is understanding the star formation process. Star formation is an enormously difficult problem because it encompasses gravity, hydrodynamics, radiative transfer, and magnetic fields. Further, the difference in density between between the initial gas cloud from which the star forms and the final star itself is 21 orders of magnitude, or equivalently, a change in physical scale of 7 orders of magnitude.

One of the most significant questions in this field is: Why do most stars form in binary systems? To address this question, numerical simulations have been run that follow the fragmentation of a large cloud of gas. The methods used have been primarily Lagrangian ones (such as SPH), although Adaptive Mesh Refinement (AMR) techniques [Berger and Collela 1989] are becoming more popular. The reason for the growth of interest in AMR is that recent results have demonstrated a severe error in a large body of numerical simulations of the cloud fragmentation process: they lacked resolution to adequately follow the balance between gravitational forces and local pressure forces [Truelove et al. 1997]. Simulations currently suggest that turbulent fragmentation plays a critical role in determining the formation of multiple star systems and that a filamentary structure is the main mechanism for transferring mass to the protostellar disk [Klein et al. 2000]. A similar process seems to govern formation of the first stars in the Universe [Abel et al. 2002].

Studying the interstellar medium presents different challenges. Traditionally, the ISM is understood as having a series of distinct phases that determine local star formation [McKee and Ostriker 1977], with regulation of the phases provided by heating and cooling mechanisms. Stellar winds and supernovae are the primary heating mechanism, while radiative cooling is the dominant cooling mechanism for the hot gas phases. The supernovae and winds also constantly stir the ISM, which in combination with rapid radiative cooling, serve to make it a highly turbulent medium. Turbulent media are difficult to understand because motions on very large scales can quickly couple to motions on much smaller scales,

and thus accurate modeling requires resolution of large and small scales [Mac Low 2000]. Because of this range of scales, achieving sufficient resolution to be able to accurately model turbulence is difficult, and a number of researchers rely on two-dimensional models to provide sufficient dynamic range. Recent simulations have shown that self-gravity alone, without the stirring provided from supernovae explosions, is sufficient to produce the spectrum of perturbations expected from analytical descriptions of turbulence [Wada et al. 2002]. In the near future, three-dimensional simulations with a similar resolution to two-dimensional models will be possible, although the incorporation of MHD turbulence makes large-scale three-dimensional simulations a formidable challenge.

33.4.8 Cosmology

The study of the Big Bang and quantum gravity epoch is still largely conducted analytically, although some aspects of this research lend themselves to computer algebra. Following these earliest moments, the Universe undergoes a series of phase transitions (or "symmetry breaking") as the forces of nature separate out of the "Unified Field" [Kolb and Turner 1993]. Computation has been used to examine the nature of the phase transitions that occur as each of the forces separates. For example, the Electroweak phase transition has been extensively examined using lattice calculations to explore whether the phase transition is first (most probable) or second order [Kajantie et al. 1993]. Numerical simulations have also investigated how non-uniform symmetry breaking can lead to the formation of defects [Ach'ucarro et al. 1999].

Computation is used extensively in the study of **Big Bang Nucleosynthesis** (BBN) and the relic Cosmic Microwave Background (CMB). However, at present, CMB data analysis probably represents the biggest challenge computationally. Theoretical modeling of BBN dates back to the 1940s [Alpher et al. 1948], and a very detailed numerical approach to solving the coupled set of equations describing the reaction network was developed comparatively early [Wagoner et al. 1967]. Currently, there are a number of BBN codes available, and considerable effort has been put into reconciling results from different codes. CMB modeling is comparatively straightforward because the equations describing the evolution of a thermal spectrum of radiation in an expanding Universe are not overly complex. However, because the CMB spectrum we measure has foreground effects superimposed upon it (such as clusters of galaxies), a large amount of effort is expended simulating the effect of foreground pollution [Bond et al. 2002].

The theoretical modeling of large-scale structure in the Universe has relied heavily on computation. Because on large-scales "dark matter" dominates dynamics, only gravity need be included, and a Newtonian approximation can be used without significant error. Particle-based algorithms are used to evolve an initially smooth distribution of particles into a clustered state representative of the Universe at its current epoch. The first simulations with moderate resolution (3×10^5 mass elements) of the distribution of galaxies were conducted in the early 1980s [Efstathiou and Eastwood 1981]. Simulations have played a leading role in establishing the accuracy of the **Cold Dark Matter** (CDM) model of structure formation [Blumenthal et al. 1984]. In this cosmological model, structures are formed via a hierarchical merging process. Simulation has also shown that dark matter tends to form cuspy halos that have a universal core profile [Navarro et al. 1997], while the large-scale distribution of matter is dominated by filamentary structures.

33.4.9 Galaxy Clusters, Galaxy Formation, and Galactic Dynamics

The modeling of clusters of galaxies and galaxy formation relies on the same codes as the study of large-scale structure, with the addition of hydrodynamics to model the gas that condenses to form stars and nebulae within galaxies. Typically, the hydrodynamic methods used are either Eulerian grid-based algorithms or Lagrangian particle-based methods [Frenk et al. 1999], although, as in star formation, AMR methods are being adopted. Modeling of galaxy clusters is comparatively straightforward because the intracluster gas tends toward hydrostatic equilibrium. However, simulations have shown that the gas in galaxy clusters shows evidence of an epoch of preheating [Eke et al. 1998].

Galaxy formation is an exceptional difficult problem to study numerically because the evolution of the gas is strongly affected by supernovae explosions that occur on scales smaller than the best simulations

can currently simulate [White 1997]. The physics is also technically challenging because galaxy formation occurs in the very nonlinear regime of gravitational collapse while simultaneously being a radiation hydrodynamics problem (although an optically thin approximation for the gas works very well). Only within the past few years has a sufficient understanding evolved to enable simulations of galaxy formation to produce moderate facsimiles of the galaxies we observe [Thacker and Couchman 2001]. Nonetheless, the very best simulations continue to lack both important physics and sufficient resolution to describe the galaxy formation process in great detail.

The first large-scale numerical studies of galactic dynamics were conducted in the 1960s [Hockney 1967]. At least initially, and to a large extent today, most n-body simulations are used to confirm analytic solutions derived from idealized models of galactic disks [Binney and Tremaine 1987]. Typically, these simulations begin with a model of a given galaxy, which usually consists of a disk of stars and an extended dark halo, which is then perturbed in some fashion to mimic the phenomenon under study. Recently, it has been highlighted that accurate modeling of galactic dynamics in CDM universes is exceptionally difficult due to coupling between the substructure in the larger galactic halo and the galactic disk [Weinberg and Katz 2002]. Traditionally, researchers believed that approximately 1 million particles were sufficient to model galaxies reasonably; however, these new results have pushed that estimate at least an order of magnitude higher. Although researchers in this field use codes similar to those in large-scale structure, specialist codes, which are designed to reduce numerical noise, have been developed (e.g., the Self Consistent Field code of Hernquist and Ostriker [1992]).

33.4.10 Numerical Relativity

General relativity calculations are extremely computationally demanding because not only is the theory exceptionally nonlinear, but there are a number of elliptic constraint equations that must be satisfied at each iteration. Further, the rapid change of scales that can accompany collapse problems often requires adaptive methods to resolve. There are also other subtleties related to the boundary conditions around black holes that present severe intellectual challenges. Currently, the strongest science driver behind these calculations is the need to calculate the gravitational wave signal of cataclysmic events (such as binary black hole coalescence), which may be detectable with the LIGO gravitational wave detector (http://www.ligo.caltech.edu/). Because of the large amount of computation involved in computing space-time geometry, and the comparatively low amount of communication between processors, numerical relativity is an ideal candidate for Grid-based computation. The CACTUS framework has been developed to aid such calculations (http://www.cactuscode.org).

Relativity calculations are most often mesh based (although spectral methods are used occasionally). Before determining the evolution equations for the space-time, a gauge must first be decided upon, and the most common gauge is the so called "$3+1$ formalism," where the space-time is sliced into a one-parameter family of space-like slices. Other formulations exist, such as the characteristic formalism, and in general the gauge is chosen to suit the problem being studied. Initial conditions for the space-time are provided and then the simulation is integrated forward, with suitable boundary conditions being applied. Building upon this body of research, the first calculation of the gravitational waveform from binary black-hole coalescence was performed in 2001 [Alcubierre et al. 2001].

33.4.11 Compact Objects

The study of compact objects such as white dwarf and neutron stars presents a formidable theoretical challenge. These systems exhibit extreme density, in turn requiring detailed nuclear physics as well as relativistic descriptions. Compact objects are widely believed to be the source of energy behind GRBs, with energetic scenarios, such as sudden mergers, driving a highly relativistic "fireball" shock wave that produces an extreme amount of gamma-ray radiation during collisions with other shock waves or the interstellar medium [van Putten 2001].

Neutron star collisions have been simulated for similar reasons to black holes: the calculation of their gravitational wave spectrum [Calder and Yang 2002]. Neutron stars are also the beginning point of core collapse (Type II) supernovae, and the simulation of the ignition process has attracted much attention. Fully general relativistic models are now beginning to appear [Bruenn et al. 2001]. Of particular interest is how the neutrinos drive a wind shortly after collapse begins [Burrows et al. 1995].

Type I supernovae occur when white dwarfs accrete sufficient mass to exceed the Chandrasekhar limit and subsequently undergo collapse. The physics is challenging because the process occurs far from equilibrium and entails radiative transfer as well as hydrodynamic instabilities. As in studies of Type II supernovae, to date most calculations use a two-dimensional approximation [Niemeyer et al. 1996] and Eulerian approaches; however, some explorations have used SPH in three-dimensions [Bravo and Garcia-Senz 1995]. The push toward full, high-resolution three-dimensional calculations is gathering momentum [Reinecke et al. 2002]. Such simulations are necessary to fully understand instabilities and include more accurate physics. However, the computational challenge is significant and, ultimately, progress awaits the development of 100-Teraflop computers.

33.4.12 Parallel Computation in Astrophysics

Parallel computing in astrophysics is often used to examine problems that simply cannot be addressed on a desktop computer, regardless of how long one could wait. The primary driver in this case is the large amount of memory available in parallel computers as compared to serial ones: the largest parallel computations simply do not fit into a desktop. The secondary use of parallel computation is to speed up data analysis, which involves performing the same analysis on many subsets of data. In this case, parallel computers significantly help in lowering the data reduction time.

More than 20% of the total cycles at the U.S. National Center for Supercomputing Applications are devoted to astrophysics, which is second only to materials science in terms of resource usage. Astrophysicists have a history of developing unique and ingenious parallel algorithms to solve the problems they face. Indeed, a number of research problems in astrophysics, such as binary black-hole coalescence and the formation of galaxies, are considered to be computational "Grand Challenges" by the National Science Foundation. These problems have computing demands that are similar to the nuclear ignition codes in the Accelerated Strategic Computing Initiative, which is part of the U.S. Government Stockpile Stewardship Program.

Parallel codes for distributed memory platforms are most often developed using the Message Passing Interface (MPI). Prior to the standardization of MPI, the Parallel Virtual Machine (PVM) API was very popular, and PVM remains the most common mechanism for parallelizing simple codes. Higher-performance APIs, such as the remote direct memory access provided provided by MPI-2, are yet to receive significant attention, primarily because vendors have failed to provide full support for this emerging standard. All of these APIs can lead to a significant increase in the size of a parallel program compared to the serial one. It is not uncommon for parallel codes to be over twice the length of serial ones. Codes written using these APIs have the potential to scale to many hundreds of processors.

Shared memory parallel codes are most often developed using the standardized OpenMP API. This API is particularly simple to use because it enables simple parallelization of codes using "pragmas" that are inserted into the code before iterative loops. The iterations within the loop, provided that they meet certain data independence requirements, can then be distributed to different CPUs, thereby speeding up execution times. The OpenMP API often does not lead to significantly longer parallel programs, but is limited in terms of scalability by the requirement of running on shared memory computers, which typically have a maximum of around 32 processors.

Over the past few years it has become increasingly apparent that although the physics being simulated by two codes is often quite different, the underlying data structures being used, such as grids or trees, are quite similar. This has led to the development of skeleton packages in which researchers need only add the numerical implementation of their equations and the communication between processors is handled by the package. CACTUS and PARAMESH are examples of this type of framework. However, most researchers seem reluctant to rely on these packages and instead develop an optimized communication layers themselves.

It is unclear what role "The Grid" will play in the development of theoretical modeling. Grid technologies have an inherently large latency, which makes simulation of elliptic-like problems (where the solution at one point depends on all others) particularly difficult. Relativistic problems appear to be comparatively well-suited to a grid environment due to the exceptionally large amount of computation involved in tensor calculations. So far, demonstration calculations run on a Grid environment using CACTUS are the best example of an effective Grid application. The potential of The Grid in astrophysics will probably be realized through extensive data analysis (akin to the SETI@home model) and data mining.

33.5 Research Issues and Summary

In this section we first highlight a number of key areas in computational astrophysics that will receive increasing attention in the future. No significance should be attached to the ordering of the items.

1. *Development of a Virtual Observatory.* A Virtual Observatory is a geographically distributed, Web-based repository of space-based and ground-based digital sky surveys, observatory and mission archives, and astronomy data and literature services. Virtual Observatories will provide powerful data analysis, visualization, and data mining toolkits for the effective and rapid scientific exploration of the resulting massive data sets. Virtual Observatories are to be enabled by technology, but driven by science.
2. *Planet formation.* Models need to include both hydrodynamics and dust to better understand the formation process in both the inner and outer solar system. Increasing resolution and space-borne optical interferometry will require detailed theoretical support from simulations to analyze observations.
3. *Star formation.* Following the full collapse of a gas cloud through to nuclear ignition remains an enormously challenging theoretical proposition. There are also a large number of questions about the formation of stars in groups, especially globular clusters, that will require key insights from simulations.
4. *Galaxy formation.* Progress is limited by both a lack of numerical resolution and an incomplete understanding of the physics of the process. It is currently unclear how to model the extremely important effect of heating from supernovae on protogalactic collapse. New discoveries will depend heavily on the growth of understanding of the star formation process.
5. *CMB data analysis.* New approximation algorithms are required to reduce the enormity of the analysis calculations for high-resolution data sets. Brute force and larger computers are not the answer, as even Petaflop computers will take over a year to analyze upcoming data sets.
6. *Black-hole coalescence.* With the growth of gravitational wave astronomy, calculating wave forms for binary systems will become increasingly important. These calculations have the potential to exploit the growth of "The Grid" computing environment.
7. *Supernovae ignition.* Current models can only simulate with adequate resolution in two-dimensions. A 100-fold improvement in computational power is needed to enable full three-dimensional calculations.
8. *Visualization.* The increasing complexity of data sets makes analysis more and more difficult. Haptic interfaces and/or multidimensional representation will be necessary to understand the phenomenological aspects of the next generation of astronomical data sets and simulations.
9. *Data mining tools that incorporate measurement error*. Efforts are being undertaken to include measurement errors in the performance of data mining tools. Pattern recognition algorithms currently define classes without regard to data quality; poorly measured data can smear out true clusters that would be easily visible if the data were limited to that of high quality. In some cases, poor-quality measurements can cause data to artificially cluster (e.g., when an upper flux limit is used in lieu of a measurement). Because the scientific method is statistical in nature, computational tools for science are needed that better reflect statistical uncertainties.

10. *Autonomous spacecraft/instrumentation operation.* Autonomous spacecraft are being designed to carry out day-to-day operations independent of ground control. Some of these operations include navigation and the scheduling and execution of observations and experiments. Support technologies for autonomous spacecraft include robotics, artificial intelligence, and control theory. This approach will reduce mission operation costs while simultaneously allowing orbital satellites to work dynamically rather than passively.

Astrophysics is gradually and irreversibly becoming a computational science. Astronomical data are being stored in and retrieved from progressively larger databases; data and metadata are being used by larger audiences. The analyses of such data are aided by pattern recognition algorithms and improved data visualization tools. Theoretical modeling has developed beyond the point where elegant calculations using relatively simple assumptions suffice; detailed models with many physical parameters are often required to adequately explain the detailed observations made by the newest orbital and ground-based telescopes spanning the electromagnetic spectrum. Parallel computation is often used in carrying out time-consuming calculations. High precision is needed to accurately calculate model parameters, and computationally intensive statistical tools are required to evaluate theoretical model efficacies. Astrophysics is thus evolving, and the new generation of astrophysicist is increasingly well-versed in the use of computational tools. This can only help us better understand the structure, evolution, and nature of the universe.

Defining Terms

Adaptive optics: An active form of observing light (rather than a passive one) in which lasers are fired at the sky from the telescope site, and the reflected laser light is used to differentially correct observed starlight for atmospheric refraction and convection.

Astrophysics: The study of the physical properties, structure, kinematics, dynamics, and evolution of celestial objects.

Big Bang nucleosynthesis: The epoch of the early Universe when the nuclei of atoms were formed. Theories of this epoch reproduce observations of elemental abundances exceptionally well, strongly supporting the Hot Big Bang theory.

Cold dark matter: Hypothesized sub-atomic particles that contribute roughly 30% of the mass in the Universe. Although in the early Universe these particles interacted vigourously, their influence is now seen only via gravitation. Theories that include the effects of these particles match observations exceptionally accurately.

Cosmic microwave background: Relic electromagnetic radiation left over from 300,000 years after the Big Bang. The distribution of radiation is very close to being homogeneous and isotropic, but carries small perturbations on it that correspond to perturbations of the matter density in the early Universe.

Cosmology: The branch of astrophysics dealing with the structure and evolution of the Universe on the largest scales.

Electromagnetic spectrum: Classification of light by wavelength (which is inversely related to energy). Ranging from light with the shortest wavelength (and largest energy) to that with the longest wavelength (and smallest energy), the electromagnetic spectrum includes gamma-ray, x-ray, ultraviolet, visible, infrared, and radio.

Galaxy: A large, gravitationally bound ensemble of stars, gas, and dust. Galaxies are traditionally classified as spiral, elliptical, irregular, or peculiar, based on their morphological structure. Our *Milky Way* galaxy is a large spiral galaxy containing more than a hundred billion stars.

General relativity: Einstein's theory of gravitation, which casts gravity not as a force, but rather as a consequence of warping in the local space-time continuum. The theory relies strongly upon geometrical concepts of distance and curvature.

Interstellar medium (ISM): The material between the stars. It is composed primarily of hydrogen and helium gas, with 1 to 2% heavy elements chemically mixed to form larger molecules ("dust"). Dense, cool molecular clouds are the seeds of star formation.

Passband: An electromagnetic regime defined by the spectral response of a particular filter and/or instrument.

Plasma: An ionized gas. A plasma behaves differently than a normal gas (which can often be modeled as a fluid) because a treatment of electromagnetic theory is needed to address the electrical charges found within it.

Supernova: A massive stellar explosion during which a star can briefly brighten to almost a billion times its original luminosity. Such events can be seen at enormous distances, although maximimum luminosity usually lasts only for tens of days.

Virtual observatory: Proposed publicly accessible metadatabase of archived ground-based, balloon, and satellite astronomical observations. The database will also be associated with a variety of search engines and data mining tools.

References

Abel, T., Bryan G., and Norman, M. L. 2002. The Formation of the First Star in the Universe. *Science* 295:93–98.

Abry, P. Goncalves, P., and Flandrin, P. 1995. *Lect. Notes in Statistics 103, Wavelets and Statistics* 15, A. Antoniadis and G. Oppenheim, Editors (New York: Springer).

Alcubierre, M. et al. 2001. 3-D Grazing Collision of Two Black Holes. *Phys. Rev. Lett.* 87:271103.

Alpher R., Bethe, H., and Gamov, G., 1948. The Origin of Chemical Elements. *Phys. Rev.* 73:803–804.

Ach'ucarro, A., Borrill, J., and Liddle, A. R. 1999. The Formation Rate of Semilocal Strings. *Phys. Rev. Lett.* 82:3742–3749.

Barnes, J. and Hut, P., 1986. A Hierarchical O(NlogN) Force-Calculation Algorithm. *Nature* 324:446–449.

Barthelmy, S., Cline, T., and Butterworth, P. 2001. GRB Coordinates Network (GCN): A Status Report, *Gamma 2001: Gamma-Ray Astrophysics: AIP Conference Proceedings, Vol. 587*, 213; S. Ritz, N. Gehrels, and C. R. Shrader., Editors (Melville, NY: American Institute of Physics).

Beckers, J. M. 1993. Adaptive Optics for Astronomy — Principles, Performance, and Applications. *Annu. Rev. Astron. Astrophys.* 31:13–62.

Beer, M. E. and Podsiadlowski, P. A. 2002. General Three-Dimensional Fluid Dynamics Code for Stars in Binary Systems. *Mon. Not. Roy. Astron. Soc.* 335(2):358–368.

Binney, J. and Tremaine, S., 1987. Galactic Dynamics. (Princeton Unviersity Press: Princeton NJ).

Blumenthal, G. R., Faber, S. M., Primack, J. R., and Rees M. J. 1984. Formation of Galaxies and Large-Scale Structure with Cold Dark Matter. *Nature* 311:517–525.

Bond, J. R. et al. 2002. The Sunyaev-Zeldovich Effect in CMB-Calibrated Theories Applied to the Cosmic Background Imager Anisotropy Power at $l > 2000$. http://arxiv.org/abs/astro-ph/0205386.

Borrill, J. D., 1999. MADCAP — The Microwave Anisotropy Dataset Computational Analysis Package. *Proceedings of the 5th European SGI/Cray MPP Workshop.* http://xxx.lanl.gov/abs/astro-ph/9911389.

Boyce, P. B., Tenopir, C., and Milkey, R. W. 2001. Electronic Journal Usage Patterns in Astronomy. *Bull. Am. Astron. Soc.* 199:1004B.

Bravo, E. and Garcia-Senz, D., 1995. Smooth Particle Hydronamics Simulations of Deflagrations in Supernovae. *Astrophys. J.* 450:L17–L21.

Brown, T. M., Charbonneau, D., Gilliland, R. L., Noyes, R. W., and Burrows, A. 2001. Hubble Space Telescope Time-Series Photometry of the Transiting Planet of HD 209458. *Astrophys. J.* 552:699.

Bruenn, S. W., DeNisco, K. R., and Mezzacappa, A. 2001. General Relativistic Effects in the Core Collapse Supernova Mechanism. *Astrophys. J.* 560:326–338.

Buchler, J. R., Kollath, Z., and Marom, A. 1997. An Adaptive Code for Radial Stellar Model Pulsations. *Astrophys. Space Sci.* 253(1):139–160.

Burrows, A. Hayes, J., and Fryxell, B., 1995. On the Nature of Core-Collapse Supernova Explosions. *Astrophys. J.* 450:830–850.

Buzasi, D. Catanzarite, J., Laher, R., Conrow, T., Shupe, D., Gautier, T. N., III, Kreidl, T., and Everett, D. 2000. The Detection of Multimodal Oscillations on Alpha Ursae Majoris. *Astrophys. J.* 532:133.

Calder, A. C. and Yang, E. Y. M. Numerical Models of Binary Neutron Star System Mergers. II. Coalescing Models with Post-Newtonian Radiation Reaction Forces. *Astrophys. J.* 570:303–313.

Clarke, D. A. and West, M. J. 1997. *ASP Conf. Ser. 123: Computational Astrophysics; 12th Kingston Meeting on Theoretical Astrophysics.*

Collela, P. and Woodward, P. 1984. The Piecewise Parabolic Method for Gasdynamical Simulation. *J. Comp. Phys.* 54:174.

Condon, J. J. et al. 1998. The NRAO VLA Sky Survey. *Astron. J.* 115(5):1693–1716.

Córsico, A. H. and Benvenuto, O. G. 2002. A New Code for Nonradial Stellar Pulsations and its Application to Low-Mass, Helium White Dwarfs. *Astrophys. Space Sci.* 279(3):281–300.

Cox, A. N., Bowers, D. L., and Brownlee, R. R. 1960. A Method of Computing Stellar Interior Models. *Astron. J.* 65:487.

Crane, P. C. 2001. Applications of the DFT/CLEAN Technique to Solar Time Series. *Solar Phys.* 203:381.

Djorgovski, S.G. et al. 2002. The Digital Palomar Observatory Sky Survey (DPOSS): General Description and the Public Data Release. *Bull. Am. Astron. Soc.* 200:6006D.

Efstathiou, E. and Eastwood, J. W. 1981. On the Clustering of Particles in an Expanding Universe. *Mon. Not. R. Astr. Soc.* 194:503–525.

Eke, V. R., Navarro, J. F., and Frenk, C. S. 1998. The Evolution of X-Ray Clusters in a Low-Density Universe. *Astrophys. J.* 503:569–592.

Evrard A. E. et al. 2002. Galaxy Clusters in Hubble Volume Simulations: Cosmological Constraints from Sky Survey Populations. *Astrophys. J.* 573:7–36.

Falgarone, E. and Passot, T. 2003. Turbulence and Magnetic Fields in Astrophysics. *Lecture Notes in Physics, Vol. 614* (Spinger-Verlag: Heidelberg).

Galama, T. J. 1999. The Effect of Magnetic Fields on Gamma-Ray Bursts Inferred from Multi-Wavelength Observations of the Burst of 23 January 1999. *Nature* 398:394.

Gilliland, R. L., Bono, G., Edmonds, P. D., Caputo, F., Cassisi, S., Petro, L. D., Saha, A., and Shara, M. M. 1998. Oscillating Blue Stragglers in the Core of 47 Tucanae. *Astrophys. J.* 507:818.

Gingold, R. A. and Monaghan, J. J., 1977. Smoothed Particle Hydrodynamics — Theory and Applications to Non-Spherical Stars. *Mon. Not. R. Astr. Soc.* 204:715–733.

Ginsparg, P. 1996. Winners and Losers in the Global Research Village. Conference held at UNESCO HQ, Paris. http://arXiv.org/blurb/pg96unesco.html.

Hakkila, J., Myers, J. M., Stidham, B. J., and Hartmann, D. H. 1997. A Computerized Model of Large-Scale Visual Interstellar Extinction. *Astron. J.* 114:2043.

Hakkila, J., Giblin, T. W., Roiger, R. J., Haglin, D. J., Paciesas, W. S., and Meegan, C. A. 2003. How Sample Completeness Affects Gamma-Ray Burst Classification. *Astrophys. J.* 582:320–329.

Hanisch, R. 1999. Electronic Preprints. *Bull. Am. Astron. Soc.* 31:1519.

Hanisch, R. J., Farris, A., Greisen, E. W., Pence, W. D., Schlesinger, B. M., Teuben, P. J., Thompson, R. W., and Warnock, A., III 2001. Definition of the Flexible Image Transport System (FITS). *Astron. Astrophys.* 376:359–380.

Heikkila, C. W., McGlynn, T. A., and White, N. E. 1999. Astrobrowse: A Web Agent for Querying Astronomical Databases. *ASP Conf. Ser. 172: Astronomical Data Analysis Software and Systems VIII.* 8:221. D. M. Mehringer, R. L. Plante, and D. A. Roberts, Editors.

Henyey, L. G., Lelevier, R., and Levee, R. D. 1959. Evolution of Main-Sequence Stars. *Astrophys. J.* 129:2.

Hernquist, L. and Ostriker, J. P. 1992. A Self-Consistent Field Method for Galactic Dynamics. *Astrophys. J.* 386:375–397.

Hockney, R. W., 1967. Gravitational Experiments with a Cylindrical Galaxy. *Astrophys. J.* 150:797–806.

Hockney, R. W. and Eastwood, J. W., 1981. Computer Simulation Using Particles. (New York: McGraw-Hill).

Kajantie, K., Rummukainen, K., and Shaposhnikov, M. 1993. A Lattice Monte Carlo Study of the Hot Electroweak Phase Transition. *Nucl. Phys.* B407:356–372.

Kiss, L. L. Szabó, G. M., Sziládi, K., Furész, G., Sárneczky, K., and Csák, B. 2001. A Variable Star Survey of the Open Cluster M37. *Astron. Astrophys.* 376:561–567.

Klein, R. I., Fisher, R., and McKee C. F. 2001. The Formation of Binary Stars. *Proceedings of IAU Symp. 2001*, held 10–15 April 2000, in Potsdam, Germany ASP. H. Zinnecker and R.D. Mathieu. editors.

Kleinmann, S.G. et al. 1994. The Two Micron All Sky Survey. *Experimental Astronomy* 3:65–72.

Kolb, E. W. and Turner, M. S. 1993. The Early Universe. (Perseus Publishing: Cambridge).

Kurtz, M. J. et al. 2000. The NASA Astrophysics Data System: Overview. *Astron. Astrophys. Suppl.* 143:41–59.

Kurucz, R.L. 1969. A Matrix Method for Calculating the Source Function, Mean Intensity, and Flux in a Model Atmosphere. *Astrophys. J.* 156:235–240.

Lasenby, A., Barreiro, B., and Hobson, M. 2001. Regularization and Inverse Problems. *Mining the Sky, Proceedings of the MPA/ESO/MPE Workshop*, held at Garching, Germany, 31 July–4 August, 2000, 15. A. J. Banday, S. Zaroubi, and M. Bartelmann, Editors (Heidelberg: Springer-Verlag).

Louys, M., Starck, J. L., Mei, S., Bonnarel, F., and Murtagh, F. 1999. Astronomical Image Compression. *Astron. Astrophys. Suppl.* 136:579–590.

Mac Low, M.-M. 2000. The Dynamical Interstellar Medium: Insights from Numerical Models. *Stars, Gas, and Dust in Galaxies ASP Conference Series: San Francisco* D. Alloin, K. Olsen, and G. Galez, Editors.

Mayer, L., Quinn, T., Wadsley, J., and Stadel, J. 2002. Formation of Giant Planets by Fragmentation of Protoplanetary Disks. *Science* 298:1756.

Mazzarella, J. M., Madore, B. F., and Helou, G. 2001. Capabilities of the NASA/IPAC Extragalactic Database in the Era of a Global Virtual Observatory. *Proc. SPIE 4477: Astronomical Data Analysis.* 20–34, J.-L. Starck and F. D. Murtagh, Editors.

McKee, C. F. and Ostriker, J. P. 1977. A Theory of the Interstellar Medium — Three Components Regulated by Supernova Explosions in an Inhomogeneous Substrate. *Astrophys. J.* 218:148–169.

McGlynn, T., Scollick, K., and White, N. 1998. SKYVIEW: The Multi-Wavelength Sky on the Internet. *New Horizons from Multi-Wavelength Sky Surveys, Proceedings of the 179th Symposium of the International Astronomical Union*, held in Baltimore, USA August 26–30, 1996, 465. Kluwer Academic Publishers, B. J. McLean, D. A. Golombek, J. J. E. Hayes, and H. E. Payne, Editors.

Meynet, G. and Maeder, A. 2002. Stellar Evolution with Rotation. VIII. Models at $Z = 10^{-5}$ and CNO Yields for Early Galactic Evolution. *Astron. Astrophys.* 390:561–583.

Murray, J. and Holman, M., 1999. The Origin of Chaos in the Outer Solar System. *Science* 283:1877.

Ochsenbein, F., Bauer, P., and Marcout, J. 2000. The VizieR Database of Astronomical Catalogues. *Astron. Astrophys. Suppl.* 143:23–32.

Navarro, J. F., Frenk. C. S., and White, S. D. M. 1997. A Universal Density Profile from Hierarchical Clustering. *Astrophys. J.* 490:493–508.

Niemeyer, J. C., Hillebrandt, W., and Woosley, S. E., 1996. Off-Center Deflagrations in Chandrasekhar Mass Type IA Supernova Models. *Astrophys. J.* 471:903–914.

Nikolaev, S., Weinberg, M. D., Skrutskie, M. F., Cutri, R. M., Wheelock, S. L., Gizis, J. E., and Howard, E. M. 2000. A Global Photometric Analysis of 2MASS Calibration Data. *Astron. J.* 120(6):3340–3350.

Paciesas, W. S. et al. 1999. The Fourth BATSE Gamma-Ray Burst Catalog (Revised). *Astrophys. J. Suppl.* 122:465–495.

Peraiah, A. 2001. An Introduction to Radiative Transfer. (Cambridge University Press: Cambridge).

Piña, R. K. and Puetter, R. C. 1993. Bayesian Image Reconstruction — The Pixon and Optimal Image Modeling. *Publ. Astron. Soc. Pac.* 105(688):630–637.

Piro, L. et al. 2002. The Bright Gamma-Ray Burst of 2000 February 10: A Case Study of an Optically Dark Gamma-Ray Burst. *Astrophys. J.* 577:680–690.

Poretti, E., Buzasi, D., Laher, R., Catanzarite, J., and Conrow, T. 2002. Asteroseismology from Space: The Delta Scuti Star Theta 2 Tauri Monitored by the WIRE Satellite. *Astron. Astrophys.* 382:157–163.

Pronik, I. I., Merkulova, N. I., and Metik, L. P. 1999. Characteristics of the Variability of the Nucleus of NGC 1275 in the Optical in 1982–1994. *Astron. Astrophys.* 351:21–30.

Ransom, S. M., Eikenberry, S. S., and Middleditch, J. 2002. Fourier Techniques for Very Long Astrophysical Time-Series Analysis. *Astron. J.* 124:1788–1809.

Reinecke, M., Hillebrandt., W., and Neimeyer, J. 2002. Three-dimensional Simulations of Type Ia Supernovae. *Astron. Astrophys.* 391:1167–1172.

Richardson, D. C., Quinn, T., Stadel, J., Lake, G. 2000. Direct Large-Scale N-Body Simulations of Planetesimal Dynamics. *Icarus* 143:45–59.

Roberts, D. H., Lehar, J., and Dreher, J. W. 1987. Time Series Analysis with Clean — Part One — Derivation of a Spectrum. *Astron. J.* 93:968.

Roden, J., Burl, M. C., and Fowlkes, C. 1999, The Diamond Eye Image Mining System. In *Demo for the Scientific and Statistical Database, Management Conf.,* (Cleveland, OH), June 1999.

Scargle, J. D. 1982. Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data. *Astrophys. J.* 263:835.

Scargle, J. 1997. Astronomical Time Series Analysis: New Methods for Studying Periodic and Aperiodic Systems. *Applications of Time Series Analysis in Astronomy and Metrology*, 215. (London: Chapman and Hall).

Scargle, J. 1998. Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data. *Astrophys. J.* 504:405.

Schou, J. and Buzasi, D. L. 2001. Observations of P-modes in Alpha; Cen. *Proceedings of the SOHO 10/GONG 2000 Workshop: Helio- and Asteroseismology at the Dawn of the Millennium*, EAS SP-464:391, A. Wilson, Ed., (Noordwijk: ESA Publications Division).

Skilling, J. 1998. Massive Inference and Maximum Entropy. Maximum Entropy and Bayesian Methods. *Proceedings of the 17th International Workshop on Maxiumum Entropy and Bayesian Methods of Statistical Analysis*, held in Boise, Idaho, 1997, 1. G. J. Erickson, J. T. Rychert, and C. R. Smith, Editors (Dordrecht/Boston/London:Kluwer).

Smoot, G. F. et al. 1992. Structure in the COBE Differential Microwave Radiometer First-Year Maps. *Astrophys. J.* 396:L1–L5.

Thacker, R. J. and Couchman, H. M. P. 2001. Star Formation, Supernova Feedback, and the Angular Momentum Problem in Numerical Cold Dark Matter Cosmogony: Halfway There? *Astrophys. J.* 555:L17–L20.

Torgersen, T. C. and Tyler, D. W. 2002. Practical Considerations in Restoring Images from Phase-Diverse Speckle Data. *Pub. Astron. Soc. Pac.* 114(796):671–685.

van Putten, M. H. P. M. 2001. Gamma-Ray Bursts: LIGO/VIRGO Sources of Gravitational Radiation. *Phys. Rep.* 345:1–59.

Wada, K., Meurer, G., and Norman, C. A. 2002. Gravity-driven Turbulence in Galactic Disks. *Astrophys. J.* 577:197–205.

Wade, G. A., Bagnulo, S., Kochukhov, O., Landstreet, J. D., Piskunov, N., and Stift, M. J. 2001. LTE Spectrum Synthesis in Magnetic Stellar Atmospheres. The Interagreement of Three Independent Polarised Radiative Transfer Codes. *Astron. Astrophys.* 374:265–279.

Wagoner, R. V., Fowler, W. A., and Hoyle, F., 1967. On the Synthesis of the Elements at Very High Temperatures. *Astrophys. J.* 148:3–50.

Weinberg, M. D. and Katz, N. 2002. Bar-Driven Dark Halo Evolution: A Resolution of the Cusp-Core Controversy. *Astrophys. J.* 580:627–633.

Wenger, M. et al. 2000. The SIMBAD Astronomical Database. The CDS Reference Database for Astronomical Objects. *Astron. Astrophys. Suppl.* 142:9–22.

White, S. D. M. 1997. Formation and Evolution of Galaxies. Cosmology and Large Scale Structure: Les Houches Session LX. 349–430. R. Schaeffer et al, Editors.

Woosley, S. E., Heger, A., and Weaver, T. A. 2002. *Rev. Mod. Phys.* 74:1015–1071.

York, D.G. et al. 2000. The Sloan Digital Sky Survey: Technical Summary. *Astron. J.* 120(3):1579–1587.

Zhang, Q., Fall, S. M., and Whitmore, B. C. 2001. A Multiwavelength Study of the Young Star Clusters and Interstellar Medium in the Antennae Galaxies. *Astrophys. J.* 561:727–750.

34

Computational Biology

David T. Kingsbury

Gordon and Betty Moore Foundation

34.1 Introduction

The past decade has witnessed the emergence of computational biology, in many forms, as a discipline in its own right. The application of mathematical and computational tools to all areas of biology is producing exciting results and providing insights into biological problems too complex for traditional analysis. In all areas of the life sciences, computational tools — from databases to computational models — have become commonplace in all laboratory settings. There is not a pharmaceutical or biotechnology company that does not have a computational biology or informatics group. Likewise, computational biology has emerged as an established academic field. Talent shortages and the inherent interdisciplinary nature of the field, however, have limited the rate of development of the academic sector.

The emergence of large-scale biological research efforts, such as the Drosophila and Human Genome Projects among other genome efforts, has contributed to the continued landslide of complex data. What sets biology apart from other data-rich fields is the *complexity* of its data, and the emergence of the fields of proteomics and systems biology has brought even more focus to that problem. Whereas a few years ago, biology was generally viewed as a scientific "cottage industry," with data being generated in a highly distributed mode, the recent move to large-scale projects has accelerated the generation of widely shared data. Still, biology has failed to agree on standard formats or syntax, leading to significant losses of data utility and requiring extraordinary efforts to use shared information. The new paradigm of *Discovery Science*, as contrasted with the traditional *Hypothesis-Driven* investigation, is seriously limited by the lack of widely used standards.

Thus, all areas of the biological sciences have urgent needs for the development of organized and accessible storage of biological data which incorporates the use of standardized syntax and file formats. The emergence of XML as a self-defining and common data format has been embraced by many in the life sciences. This has been accompanied by the use of **SOAP** and other tools to deploy data exchange via

Web Services. While this may be considered a form of biological database development, this approach is an attempt to enhance traditional database technology such as transaction-oriented relational database systems, which are often inadequate to serve many areas of the biological sciences. It is clear that collaboration between computer scientists and biologists continues to be necessary to design information platforms which accommodate the needs for variation in the representation of biological data, the distributed nature of the data acquisition system, the variable demands placed on different data sets, and the absence of adequate algorithms for data comparison, all of which are characteristic of biological science.

The continued advances in commercially available hardware and a greater emphasis on hardware clusters and grids are beginning to have a profound effect on computational biology. While in the past, traditional general-purpose hardware was inadequate for many of the most computationally intense problems, this condition has essentially disappeared as high-performance general-purpose instruments, commodity clusters, and grids have become more widely available. However, not only hardware limitations have affected the productivity of the computational biologist. There is a continuing need for new algorithm development to cover many tasks, especially real-time data analysis and comparisons between objects and images. Imaging technology is central to almost all of biology, and data representation through image construction remains an elusive but astoundingly powerful tool.

During the past decade there were dramatic advances in instrumentation and related methodologies for both light and electron microscopy. The advances lie not simply in higher resolution, but rather in a broader size range of structures that can be analyzed, more powerful methods for putting together the pieces of three-dimensional puzzles of cell form, and the addition of dynamic details of biological form and function, ranging from the subcellular to the physiological level. These approaches are computationally demanding. Computational resources have been expanding to meet these needs but remain inadequate for dealing with the massive data flow. As new experiential and computational approaches emerge in a few laboratories around the world, and as the ability to combine data from multiple instruments expands, the demand for better hardware and software continues. However, it is important that new software be developed within the context of the experimental research driving the needs; that is, there must be close collaboration between those developing the software and the groups carrying out research on static and dynamic structures.

X-ray crystallography and *NMR* are the major experimental methods for deducing macromolecular structures at atomic resolution. Both methods produce extremely large amounts of data and are entirely dependent upon the availability of powerful computers and sophisticated processing algorithms for the interpretation of raw data. In addition, there are fundamental scientific problems in both areas that require major computational advances. Substantial opportunities exist for combining structural information from several experimental techniques. This may provide the basis for a structural solution where only partial data is available from any single technique. With improved computational tools, combining physical data from a variety of sources has become more common. These developments will allow solutions to be obtained for structural problems which would otherwise be intractable. Analysis of errors in structures based upon experimental data from several sources also represents a significant computational challenge.

Advances in x-ray and NMR data analysis will lead directly to rapid developments in the field of protein-folding and structure-function prediction, which will be synergistic with developments in other areas of biology itself, and especially computational biology. Common problems of data representation, search strategy, pattern recognition, and data visualization appear in many fields. There is a particularly exciting synergistic relationship between the protein-folding field and the fields of structure determination by x-ray crystallography and 2-D NMR. Each field will benefit from rapid advances in the others. Improved folding algorithms provide a new way to attack the phase problem in crystallography, and new, more carefully refined protein structures provide rich new insights into protein folding.

Computational neurobiology is one of the most rapidly developing areas of computational biology. It gives us the hope of interpreting the mass of anatomical and physiological information about the nervous system, much of it derived from diagnostic testing, that is now available in functional terms. Better integration and interpretation of these data will permit neurobiology to make contact with other fields such as psychology and artificial intelligence. This work is making specific, testable predictions in the

areas of sensory perception (visual, olfactory, and auditory), memory, learning, and motor control. Above all, it will lead to the integration of all these aspects to provide an eventual understanding of the total functioning of the nervous system. Such integration can be expected to provide new insights that will lead to improvements in the treatment of diseases of the nervous system at all levels, from neuropharmacology to psychotherapy.

The area of *genome analysis* has been, and continues to be, a major focal point in computational biology, and much progress has been made over the past few years. The sequencing of the human, mouse, rat, and fruit fly genomes in the past few years has challenged computational biologists and statisticians in many areas. While robust approaches to both **linkage mapping** and **physical mapping** were developed in the past, a new set of genetic challenges emerged from the genome sequencing effort. The human genome contains a wide variety of gene variations, or *polymorphisms*, that are responsible for the unique character of each individual, including inherited disease. A single human genome contains millions of these polymorphisms, and the detection and statistically valid association of a particular polymorphism with a given trait is now a major problem in computational genomics.

In many cases this analysis requires the ability to analyze tens of thousands of markers in family pedigrees. To be fully useful in a meaningful quantitative sense, this analysis will require powerful computer simulation and modeling. Major algorithmic advances have been made in the area of sequence assembly and clone assembly in physical mapping. As biologists continue to pursue the rapid sequencing of many genomes, the most common strategy is to use "shotgun sequencing," which relies on the assembly of random fragments. While powerful assembly algorithms have been developed, many in the field still consider this an important problem.

Common to all of the problem areas mentioned is the need for good visualization of data. Visualization is necessary because the map and sequence analysis phase for a molecular biologist is equivalent to exploratory analysis for a statistician. It is at this point that the experimentalist gains the feeling for, and understanding of, a physical or linkage map or sequence, which may then guide many months of experimental work. The complexity inherent in biological systems is so great that very sophisticated methods of analysis are required. These are the tools that must be readily accessible to molecular and cellular biologists untrained in computer technology.

Ecology and evolutionary biology encompass a broad range of levels of biological organization, from the organism through the population to communities and whole ecosystems. This complexity demands computational solutions. The need for enhanced computational ability is most evident when one attempts to couple large numbers of individual units into highly interactive and largely parallel networks, whether at the tissue, community, or ecosystem level of organization. The proliferation of information from remote sensing introduces the need for geographical information systems that provide a framework for classifying information, spatial statistics for analyzing patterns, and dynamic simulation models that allow the integration of information across multiple spatial, temporal, and organizational scales.

What follows is an examination of several specific areas of computational biology, with particular emphasis on those areas related to molecular biology, and a short development of the experimental paradigm and highlights of the current computational challenges regarding the requirements for further development of that area. There are common themes that appear in several of the sections, and these themes deserve special attention because they appear to be limiting the development of the entire field, regardless of the specific area of research. (This review will not attempt to cover the important areas of computational neuroscience and ecology and evolutionary biology in any further detail. Both fields are rich in computational challenges and theory and, like some of the areas covered here, deserve a chapter of their own.)

34.2 Databases

Biology is inherently information-rich because of the complexity and variety of living systems. Understanding these systems requires information about their organization, structure, and function at a multitude of levels, from the macroscopic to the molecular. Moreover, each species (and in many species, each organism)

represents the potential for a unique solution to the problems of life processes and the organism's interaction with the environment. The full understanding of biology requires extending organismic complexity to include the relationships of species and organisms in their ecological niches, as well as the evolution of the biosphere over time.

Historically, much of this information was accessible to scientific inquiry only at high levels of abstraction, so that the inherent richness of information was not reflected in the volume of data available. This situation has changed dramatically in a number of biological fields, among them molecular biology, neurobiology, ecology, and taxonomy. This change has been made especially dramatic by the ubiquitous use of the World Wide Web (WWW). Emerging scientific paradigms will require even more data, organized into large and dynamic databases to support ongoing biological research. Indeed, some aspects of modern biology (e.g., the various genome projects, systems biology or protein structure–function studies) are now utterly dependent upon database and computer technology. In many cases, current data collections are not well organized for ease of retrieval or error correction but remain central to work in a given field. For example, several important macromolecular databases are maintained as flat files, poorly delimited and not accessible to ad hoc queries because of the absence of well-structured fields. Because of the importance of such data, investigators around the world struggle to find solutions to ready access and query. The reliance of modern molecular biology on databases is exemplified by the heavy dependence on the collection of DNA-sequence data (GenBank, EMBL, and DDBJ). These databases contain the definitive public domain DNA sequence data. However, because of the difficulty in the use of such data, and the large-scale generation of proprietary data, a new group of commercial data sets has become available (e.g., the Celera Discovery System, Incyte's LifeSeq, Gene Logic's GeneExpress, etc.).

As a result of this need, a substantial number of people now devote their careers to data management and computational analysis in biological disciplines. However, despite significant and vigorous efforts, the present generation of biological databases will fail in the next decade without significant continued development. They were not (and could not have been) designed to deal with the volume, complexity, and diversity of the data that will need to be accessible for future biological research.

The explosion of data is derived, in large part, from the desire of increasing numbers of investigators to have shared data repositories. The pressure on databases is severe in a quantitative sense, but is equally daunting in terms of the diversity and the interrelationships that must be represented among the data. These problems are further complicated by the way data is generated in biological research, which is geographically dispersed and, more importantly, lacks any meaningful standardization. Taken together, these problems pose unique transdisciplinary challenges for database design. Indeed, it is important to note that a single technology is not yet in hand that can support the design of adequate databases for much of the biology of the next decade. In fact, the application of the term "database" itself tends to be misleading. Database technology and theory as it currently exists is an inadequate paradigm for what is needed now and in the future to represent and organize biological information. For example, biological data includes mixtures of measurements, images, and interpretation, including extensive collections of metadata and derived data. Ideally, each of these data types would be available to ad hoc queries. At present, only a few extended relational systems have addressed these needs, and without complete success. The most recent focus has been to define this data in an XML format that enables the development of a variety of solutions to access and analysis, including access through Web services models.

Biologists have become increasingly sophisticated in their use of computers and in their abilities to state their research requirements in terms of informational and computational strategies. Biological science is now posing questions that not only require computational solutions, but also provide problems of fundamental interest to computer science researchers, thus creating the possibility of effective interdisciplinary work. Fortunately, there is a growing community of transdisciplinary workers whose expertise is centered at the interface of biology and computing, and who can provide much of the insight into how the two fields can interact productively.

One of the principal challenges in scientific computing in the next decade will be the development of database systems that can handle the inherent complexity of biological information and the marriage

of those systems with visualization tools that enable laboratory scientists to interpret and understand the data and analytical results. The existence and availability of such databases will transform the way the science is done, and make possible completely new paradigms of biological research. Meeting this challenge will require the construction of databases in fields where none are available, significant research and development in database and knowledge-base technology, and the provision of a robust and widely available computational infrastructure for biological science through new algorithmic approaches to data analysis and tools to embed database access into analysis and modeling.

34.2.1 Access and Communication

The emergence of the World Wide Web has revolutionized access to biological databases. Because the subdomains of biology are fragmented, it has been necessary to develop many distinct and customized databases. The fact that there is little semantic consistency between these databases has raised a significant barrier to linking them through standard query mechanisms. Several investigators have built hypertext-based linkages between a number of different databases, bringing together a richer data resource. One representative of the several available systems is the Biology Workbench developed at the San Diego Supercomputer Center and the University of California, San Diego (http://workbench.sdsc.edu/). Another approach to database linking through a common interface was developed by the European Molecular Biology Laboratory (EMBL) and has been commercialized and further refined by Lion Biosciences (http://www.lionbioscience.com/solutions/products/srs). The tool, termed SRS (sequence retrieval system), enables a scientist to use common terms to query a variety of databases and then to establsih links between them based on common features and cross references. However, it does not enable *ad hoc* SQL queries across all data sources.

As powerful as the WWW-based systems are, they still lack the potential for supporting complex *ad hoc* queries that would be achieved through a true federation of biological databases. The need for such a federation has been recognized most acutely in the genomics community, and the outlines for such a federation were developed several years ago [Robbins 1994]. It was suggested that for minimum technical linkage, all of the participating databases present similar **APIs** (application programming interfaces) to the Net. All of the databases in the federation should also be relational systems that support SQL queries. Ideally, these databases should (1) be self-documenting, (2) be stable, and (3) conform to agreed-upon federation-wide semantics. The problem with this strategy is that in many cases it places the goals of the federation in conflict with the rapidly changing nature of biological research and the needs of the specialty user community. To cope with these problems, several systems have been developed to integrate heterogeneous databases. The two most common are DiscoveryLink developed by IBM and discoveryHub developed by GeneticXchange. Both approaches utilize an intelligent broker architecture with a series of wrappers that describe the specific databases they are linking. Neither has fully solved the problem of the semantic inconsistency of biological databases; however, discoveryHub has attempted to approach this problem through a semantic translator.

34.2.2 Representation/Data Modeling

To enhance the continued development of shared data resources, we must increase database expressiveness, study representation of biological knowledge, and automate modeling of database schemas. Biological knowledge is extremely rich and diverse; it includes raw experimental data (images, numbers, symbols); interpretations of experimental data (descriptions of biological objects with complex properties and internal structures); descriptions of experimental methods (complex procedures); and theoretical knowledge (documents, equations, descriptions of processes such as gene expression). Existing database technology provides a small number of simple representational primitives (such as relations); allowing databases to capture only a small piece of this rich biological semantics. Research is needed to provide richer representational capabilities, and to study how to represent the wide range of biological knowledge. Further, because biological databases will model a large number of complex entities, biological database schemas

will be correspondingly complex. Researchers must investigate automated methods of managing database schemas to increase the efficiency of the database design process.

34.3 Imaging, Microscopy, and Tomography

Image reconstruction with light and electron microscopy and other imaging techniques is a powerful tool for the characterization of biological structures in three dimensions over a wide range of scales. Biological structures amenable to one or more techniques of imaging include single macromolecules and macromolecular assemblies, subcellular organelles, whole cells, and tissues. Characterization of the spatial organization of biological structures is critical for determining the functionality of these structures. Many fundamental problems in biology are open to study by microscopy and other imaging techniques. The advances in computer-controlled scanning microscopies, high-resolution atomic-force microscopy, and cryoelectron microscopy have substantially expanded the level of detail that can be attained by these methods.

The continuing advances and widespread applications of **transmission electron microscopy** at conventional, intermediate, and high voltages (with and without energy filtration), as well as confocal light microscopy, are leading to the production of vast amounts of data that need to be processed to extract the structural information and biologically important details. Each of these techniques can produce three-dimensional images of biologically important structures. Successes in these studies do not imply that the technical problems have been solved; currently, there remain substantial computational challenges to meet before achieving the goal of making efficient use of these techniques. These challenges include developing improved image processing and reconstruction algorithms.

In transmission electron microscopy, there have evolved three distinct methodologies for producing three-dimensional information: (1) electron crystallography of two-dimensional lattices and the processing of images of symmetric structures; (2) the analysis of multiple images of isolated asymmetric macromolecular assemblies; and (3) electron tomography, or three-dimensional reconstruction from a tilt series of images obtained from a single structure. Confocal light microscopy (CLM), a tool used in cell biology, produces three-dimensional images by scanning light focused to a single point over a three-dimensional grid in the specimen, and then imaging the light onto a point detector. Compared to conventional microscopies, CLM has a somewhat higher resolution and a much smaller depth of field, minimizing mixing of image data from different depths in the specimen.

The success of these methods has been substantial (Fernandez et al., 2002). This has stimulated the development of a comprehensive user-oriented facility, the National Center for Microscopy and Imaging Research (NCMIR). Generally known as the *Telescience Portal*, it is used by molecular and cell biologists, electron microscopists, and computer scientists around the world. The portal links instruments and computers at laboratories in Europe, North America, and Asia over a dedicated IPv6 network. NCMIR's central instrument is a state-of-the-art 400,000-volt intermediate-voltage transmission electron microscope equipped with charge-coupled detectors (CCDs) and can be operated under complete computer control and via the Internet. Because of its design, the instrument is able to penetrate thicker samples than conventional electron microscopes and is therefore particularly good at 3-D reconstruction. In collaboration with Japanese scientists, the portal also provides network-base access to a 3-MeV ultrahigh-voltage electron microscope. These resources are also linked to high-performance computing facilities in Taiwan and San Diego. Thus, the Portal is an applications environment supplying centralized access to all of the tools necessary for high-level electron tomography. A full description of the Telescience Portal and a related facility (the Biomedical Informatics Research Network (BIRN)) is available at http://www.ncmir.ucsd.edu.

Use of these imaging techniques entails three fundamental problems. The first is the vast quantity of data being produced that requires complex processing. This is a problem common to all imaging techniques within and outside biology. For example, cryoelectron-microscopic analysis of icosahedral viruses involves sample preparation followed by microscopy followed further by a series of computational steps, many of which are very computation-intensive [Cheng et al. 1995]. Because of the limited computing power available, only a limited number of independent images can be analyzed. However, three-dimensional images are intrinsically complex, and the amount of information required to characterize an image in

detail is very high. For example, construction of the three-dimensional image of a molecule from electron-microscope images of single particles routinely requires thousands of particle images. This means that the actual time to produce a three-dimensional image is weeks to months, due in large part to the user-mediated steps that remain in the analysis. For efforts like this to prove maximally fruitful, it is crucial that each step be automated as much as possible.

The second fundamental problem common to all imaging techniques is the existence of a point spread function due to the instrumental broadening that is intrinsic to each form of imaging. For example, transmission electron microscopy loses a cone of data in the Fourier transform of the image, and the restoration of this cone represents a difficult, open problem. In **scanning confocal microscopy**, the point spread function is greatly extended in the direction parallel to the optical axis, and narrowing it could improve the resulting three-dimensional images.

Third, the results of any imaging method must be quantitated and displayed. The problems of image enhancement and visualization are completely general, although each technique may benefit more from specific display modes than others. Quantitative comparison of images also provides substantial challenges, especially in the presence of noise. Comparison of two images of flexible objects (e.g., cells or chromosomes) represents a substantially greater challenge.

One of the recurring problems common to all imaging techniques is the existence of artifacts due to incomplete data collection. These artifacts may seriously interfere with the interpretation of the reconstruction, and may even lead to incorrect conclusions. This problem is most serious in electron microscopy, where a full range of viewing angles is not usually accessible, and data corresponding to a cone or wedge in Fourier space cannot be collected. In confocal microscopy, the resolution in the z-direction (parallel to the optical axis) is much lower than in the x and y directions, as reflected by a nonisotropic instrumental point spread function.

Although several restoration algorithms have been in existence for some time, only the recent increases in computational speed have made their practical implementation possible. Two algorithms with potential application to signal restoration have attracted special attention due to their success in other fields — projection onto convex sets (POCS) [Bellon and Lanzavecchia 1995] and maximum entropy (ME) [Schmeider et al. 1995]. Both methods are extremely computation-intensive, making their application to realistic-sized image volumes ($64 \times 64 \times 64$ or $128 \times 128 \times 128$) extremely demanding. The full development of these algorithms into something useful for detailed images of biological systems will require many computation cycles, to allow many different parameter values to be tested (ME) or a variety of different constraints to be used (POCS). Thus, a serious attempt to make three-dimensional image restoration viable for biological images will always require the highest available computational speed, along with advances in the theory and design of algorithms.

34.4 Determination of Structures from X-Ray Crystallography and NMR

The three-dimensional structures of proteins and nucleic acids are essential elements in the pathway relating gene sequence to function. The problem of predicting three-dimensional structure from a sequence remains unsolved at present. X-ray crystallography and NMR are the major experimental methods for deducing these structures at atomic resolution. While the rate at which these structures are being determined has increased dramatically, it lags significantly behind the rate at which new sequences are being accumulated. Each newly determined structure increases our knowledge in two ways. First, it adds to the database of known structures that can be used in knowledge-based methods in subsequent structure determinations of other macromolecules. Equally important, each new structure provides insights into the fundamental biological processes that support all life.

NMR and x-ray crystallography both produce extremely large amounts of raw data and are entirely dependent upon the availability of powerful computers and sophisticated processing algorithms for the interpretation of those data. In addition, there are fundamental scientific problems in both areas that

require major computational advances. Both NMR and crystallography make use of constraint refinement to optimize the fit of experimental data to working models, and both fields use large scale-simulations to correlate the molecular models to known biological properties. Like investigators in three-dimensional microscopy, crystallographers and NMR spectroscopists are using maximum-entropy reconstruction as a major tool in computational analysis.

34.4.1 Determination of Macromolecular Structures from NMR Data

NMR methods for determining three-dimensional structures of macromolecules in solution have become increasingly important over the past several years. Major limitations on the speed and ease of analyzing the NMR data include the difficulty of assigning individual resonances in the spectra to particular protons in the molecule and then of calculating the full three-dimensional structure using distance geometry, molecular dynamics, or algorithms. For example, when determining the structure of a 100–150-residue protein, it may take as much as two weeks of NMR spectrometer time to collect the raw data, two or three days to calculate the spectra, and months or years to fully interpret the results. The complexity of this process depends on both the size of the protein and the extent of peak overlap within the spectra. Because this is the critical bottleneck in obtaining the structural information, it limits the size of molecules that can be considered. One critical element is the solubility of biological macromolecules. The balance between solubility and the sensitivity of modern NMR equipment places the current lower limit on concentration at around 0.5 m*M*. Many proteins, especially those of high molecular mass, aggregate at such high concentrations, leading to broad spectral lines of no value in structural studies. The solution to this problem lies with the development of enhanced computational approaches, such as maximum-entropy reconstruction of specially collected data sets. This approach requires approximately 100 times the computational work of traditional discrete-Fourier-transform processing. The refinement of this approach will require the continued collaboration of structural biologists and computer scientists [Schmieder et al. 1995]. Future advances in algorithms, software development, and the availability of more powerful computers will make a major impact on the time required to interpret NMR data.

A critical step in interpreting the data is to use the relationships between protons signified by two-dimensional (or three-dimensional) crosspeaks to assign the resonances to particular protons in the molecule. Assignment is frequently the rate-limiting step in structure determination. There are a number of different strategies for assignment, and approaches to automate this process are under active investigation. It is clear that several approaches will be necessary to deal with the problems associated with ambiguous data. Both the **sequential** and **mainchain-directed** assignment procedures make use of patterns of *J*-correlated and distance-correlated relationships. In both cases, the procedure is still largely manual, although there have been recent attempts to automate parts of the analysis. The development of computer-assisted or fully automatic pattern recognition techniques would make a major impact on the time required to make the assignments, as well as the size of molecules that can be studied. This is particularly true as the complexity of the original NMR data increases.

Once protons have been assigned, a three-dimensional structure can be estimated using the peak areas from the **NOESY** spectrum to determine distance constraints. Extensive computing is required to calculate structures using distance geometry, molecular dynamics, or **Kalman filtering** techniques. Several refinements of the structure are needed, and a family of structures is usually generated. Recently, back-calculation of the NOESY spectrum has been used to try to refine the structure. The value of this procedure and the effect of using different techniques to calculate NMR structures still need to be established. However, both the development and application of this technique require major computing power.

34.4.2 X-Ray Structure Determination of Macromolecules

X-ray crystallography provides a fine example of how the availability of affordable supercomputers has dramatically sped up the rate at which x-ray structures can be determined. There are four major phases in crystal structure determination: crystal growth, solution of the phase problem, interpretation of the

resulting electron density in terms of an atomic model; and refinement of the atomic model. Up until very recently, the refinement of protein structures using x-ray data required substantial manual intervention and took one or more years to get to a satisfactory stage. With the incorporation of **simulated annealing** methods into the refinement procedure, the time required for this phase continues to decrease as more computational power becomes available. It should be noted that while the established protocols for simulated annealing can now be run fairly comfortably on workstations, the original development and testing of the method required the availability of supercomputers such as the Cray.

The fundamental problem in protein crystallography is the **phase problem**, which, with the improvement in refinement techniques, has become the major bottleneck in the structure determination process. The current method of multiple isomorphous replacement (MIR) relies on measuring data using crystals of the protein soaked in different metal compounds. The soaking procedure does not always yield suitable crystals, and in some cases the lack of metal derivatives delays the determination of the structure. Attempts to solve the phase problem without resorting to MIR or other experimental techniques fall into two classes:

1. *Ab initio phasing.* This does not rely on any additional experimental information other than that provided by the x-ray data set on the native protein crystal. Methods that have been successful for small molecules break down in the range of 100 atoms or so, well below the range of 1000 atoms or more in protein molecules. The availability of more computing power is gradually leading to the extension of some of these methods to larger polypeptides, and intensive computational effort may yield success for the smaller proteins (approximately 50 amino acids). The *ab initio* solution of virus structures may be aided by taking advantage of their high degree of noncrystallographic symmetry. If an adequate starting model is available, then these techniques allow the extension of phases directly over a wide resolution range [Rossman et al. 1985]. Novel approaches, such as those based on maximum entropy, may also provide computational tools that will have a substantial impact on obtaining phases for diffraction data.
2. *Use of prior knowledge.* In cases where the three-dimensional structure of a closely related protein is known, the phase problem can be solved using the known structure as a starting point for refinement *(molecular replacement)*. Computation-intensive methods such as simulated annealing have proved particularly valuable here, because significant distortions may have to be introduced into the starting structure and the optimization procedure needs to escape from the local minimum of the starting structure. The inclusion of prior knowledge into phase determination is likely to become extremely important in the near future, because it is becoming clear that proteins are built up from smaller subunits or segments that are commonly shared between large numbers of proteins. As the database of known structures increases, it should be possible to use segments of structures in search procedures to solve the phase problem. There is a need here to develop and apply novel search procedures and pattern recognition algorithms, many of which will also be applicable in the next stage of the structure determination.

Noisy electron density maps are difficult to interpret manually because of false or missing connections in the electron density. It is not uncommon for a structure determination project to arrive at this stage rapidly, once crystals are obtained, and then to spend a long time getting a better electron density map (the phase problem). Computing is essential in two ways. First, phase bias from the model must be minimized by optimizing relative contributions from calculated and observed diffraction data. Second, rapid interactive refinement of difficult-to-interpret regions must be performed to help to solve the problem. This requires high-speed supercomputing and a fast network link between the supercomputer and a high-performance graphics workstation.

Even in the case of a well-defined electron density map, fitting an atomic model to these maps is a time-consuming process. The problem lies in automatic recognition of characteristic patterns of macromolecular structure, such as alpha helices, that cannot be achieved by existing algorithms. Because of the complexity of the three-dimensional pattern recognition, further advances in methodology and computing speed are required.

34.5 Protein Folding

Protein folding, recognized for many years as one of biology's core problems, remains a focus of much work and attention. Folding converts the linear, one-dimensional information of the amino acid sequence into the biologically active three-dimensional structure. Folding may be thought of as a final unsolved aspect of the genetic code, and it is clear that progress on the folding problem will have tremendous theoretical and practical implications for biology. As described above, there have been dramatic advances in crystallography and two-dimensional NMR, and the virtual explosion of protein sequence data reemphasizes the importance of working on the folding problem. Even limited progress in this area could have a tremendous payoff, especially because of the recognition that such diseases as cystic fibrosis, Alzheimer's, Creutzfeldt-Jacob (human form of Mad Cow), and many others are protein folding-related problems [Dobson 1999, Fink 1998, Hammarstrm et al. 2003]. The *protein folding problem* can be considered either as:

- The problem of understanding the actual kinetic pathway by which a protein folds, or
- The problem of predicting the final folded conformation.

Obviously, a detailed structural understanding of folding intermediates would lead to a prediction of the final folded structure. However, experimental studies of folding intermediates have been very difficult because the intermediates are present in vanishingly small amounts. The rich database of known structures provides an excellent guide regarding the final folded conformation and can serve to guide theoretical work. In addition, there is an evolving literature regarding fruitful experimental approaches [Hammarström et al. 2003].

At first glance, the protein folding problem may seem to have a tantalizing simplicity (a given string of 100 amino acids contains all the data needed to determine the final folded structure), but the problem is extraordinarily complex. If each residue in a polypeptide chain can adopt ten distinct conformations, then the protein could adopt 10^{100} distinct structures, which leads to an extraordinarily difficult search problem. Many different strategies have been used in approaching the folding problem. Some methods have relied on detailed physical models of the polypeptide chain and have tried to carefully simulate the interactions (hydrogen bonds, van der Waals contacts, electrostatic interactions, etc.) that stabilize the chain. Other methods rely on the structural database that has accumulated over the past decades. In some sense, it appears that "structure is more conserved than sequence," so that the structural database is a useful guide when modeling new proteins.

Current approaches to the folding problem can generally be placed into one of two methods: direct determination of the folded confirmations, or a template-based method. Direct methods seek to determine the lowest acceptable energy point in a suitably defined conformational space. Template-based methods compare the sequence of the unknown with a collection of solved structures and select a limited number of highly scored possibilities [Luthy et al. 1992, Sali et al. 1994].

One core problem with direct methods is the difficulty of searching through the astronomical number of possible structures. A naive calculation may suggest that there are 10^{100} conformations, yet it is clear that only "a few" of these are of low enough energy to be plausible structures or plausible folding intermediates. The multiple-minima problem arises repeatedly in studies of folding. Both physical approaches (based on a detailed molecular potential surface) and pattern recognition schemes (based on analogy) encounter the same problems with multiple minima.* Stochastic search algorithms have proved especially useful in handling problems with multiple minima, and the method of **simulated annealing** is frequently used. This method corresponds to a simulation of the molecular dynamics under the influence of random thermal forces. Other search algorithms involve buildup or stochastic buildup based on genetic algorithms. To estimate the difficulty of the multiple-minima problem in protein folding, it is possible to draw upon

*The same search problem occurs in other parts of computational biology, including sequence comparison problems in genomics, neural networks, and immune-network modeling.

some parts of statistical physics for help. Simple lattice models have been used to estimate excluded-volume effects after polymer collapse, and these models indicate that the number of allowed conformations may be far smaller than suggested by the initial naive estimates. This result is very encouraging because it suggests that the search problem can focus on a much smaller region of conformational space.

Detailed atomic models have been used to study protein folding and stability. The models are based upon well-understood principles of physical chemistry and have been used in conjunction with molecular dynamics and Monte Carlo methods to study the underlying forces that determine the stability of folded proteins. The application of free-energy perturbation theory has been a particularly exciting development. These approaches are beginning to provide a much better understanding of the key forces involved in protein folding and stability, such as the true role of electrostatic interactions and the origin of the hydrophobic effect. Continued close interactions between experimental biologists and computational/theoretical researchers have also been extremely important for this field. Theory can help design new experiments, and better data can allow the refinement of basic physical models.

Although not directly linked to the protein folding problem, an extremely important area of computational biology involves the molecular modeling of protein function. A molecular understanding of enzyme catalysis can now be approached through a combination of molecular dynamics and quantum chemistry. There have been many applications of this approach, and important insights about the mechanism of triose phosphate isomerase have recently come from such simulations. Another exciting area is the recent development of computational approaches to modeling electron transfer. This a fundamental and inherently quantum-mechanical process involved in energy transduction and photosynthesis. Signal transduction is another extremely important and active area of research. This involves problems of docking and protein–protein recognition. Allosteric transitions are a frequent consequence of such interactions, and new methods for studying protein motions on long timescales are being developed [Gilson et al. 1994].

Much information has been derived from the ability to clone and express a protein of choice, followed by deliberate mutation of individual residues or larger segments. Perhaps the simplest application of folding that can be envisioned would be to predict the structural perturbation caused by a single-residue mutational change in a protein of known structure [Hammarström et al. 2003]. However, frequently, such mutations do not result in major rearrangements of the chainfold, as shown by the work of Kossiakoff [Eigenbrot et al. 1993] and others. While the structural effect of single-site changes has been successfully predicted in some cases [Desjarlais and Berg 1992], there are many conspicuous failures, and clearly more work is needed. At the level of larger segments, attempts to predict antibody hypervariable loops have also met with partial success [Tramontano and Lesk 1995; Pan et al. 1995]. Recent experiments suggest that deletion of entire loops can be tolerated while partial deletions of the same segments are not. Such results suggest the existence of quasi-independent modules, which would simplify calculations.

34.6 Genomics

Genomics is the study of DNA and its products at the genome level. This includes both experimental and theoretical aspects of the problem. For the computer scientist, the interest lies in the discrete domains and output of the system; for the biologist, the goal is to reveal the function of a sequence, either of the DNA or, more commonly, of the protein gene product. The biological effort includes genetic mapping, physical mapping (restriction maps, ordered clones, x-ray diffraction maps, cytogenetic maps, and sequence assembly), and sequence analysis including polymorphism identification and location. There is a natural division of the work into two major computational areas: support for the construction and representation of various maps, and analysis of the data produced. This overlaps at the boundary with computation involved in the analysis of protein folding and interacts with work to assign function to gene products. The systematic effort to map genomes in the presence of variability and error also mandates the use of new computational techniques to assist in the design of efficient experiments. Beyond this, there are four major areas in which the use of computers is indispensable: database searching for DNA sequence analysis, maximum-likelihood calculations for genetic linkage mapping, DNA sequence assembly, and general bookkeeping (e.g., laboratory information management systems).

34.6.1 Genetic Mapping

Genetic mapping deals with the inheritance of certain genetic **markers** within the pedigree of families. These markers can be genes, sequences associated with genetic disease (polymorphic regions, often single nucleotide changes), or arbitrary probes determined to be of significance [e.g., single nucleotide polymorphisms (SNPs), sequence tagged sites (STSs), or expressed site tags (ESTs)]. The sequence of such markers and their probabilistic distance (traditionally measured in centimorgans and now in many cases in megabase pairs) along the genome can often be determined with fair accuracy by the use of maximum-likelihood methods. Inheritance of traits across the pedigrees of multiple families is determined by a number of techniques that essentially hybridize each family member's genome against the predetermined probes. Eventually, the genetic map most likely to produce the observed data is constructed. Only a few years ago, the knowledge of the mathematics involved and the computational complexity of algorithms based on that mathematics limited analysis to no more than five or six markers. As knowledge of approximations to the formulas and likelihood estimation has improved, software capable of producing maps for 60 markers or more [Magness et al. 1993, Matise et al. 1994] has been developed. Early advances in this area include the identification of a large number of SNP and EST probes as well as software capable of tractably producing maps based on hybridization pedigree data [Cinkosky and Fickett 1993]. Further progress in this area has used mathematical methods such as combinatorics, graph theory, and statistics, and computer science methods such as search theory. Although significant progress has been made over the past few years, considerable effort is still required to make genetic linkage maps effective tools for genetic research.

The human genome contains a wide variety of gene variations that are responsible for the unique character of each individual, including inherited disease and genetic susceptibility to disease. These variations also predict a variety of responses to external factors such as chemicals and drugs. These variations are of several types and two individuals may vary by as many as 5 million locations. The association of these variations with specific traits is a significant mathematical and computational challenge. Furthermore, tools to address more complex situations (such as manic-depressive disease, which is likely to involve multiple genes) are badly needed.

34.6.2 Sequence Assembly

The recent successes in sequencing the fruit fly [Adams et al. 2000], mouse [Waterston et al. 2002] and human [Lander et al. 2001, Venter, et al. 2001] genomes were derivative of a rapid sequencing and **sequence assembly** technique that was intensely computational. The assembly technique is referred to as a "shotgun assembly" process. Shotgun assembly is an example of an inverse problem: starting with a set of sequence reads randomly sampled from a target, reconstruct the order and position of those fragments in the original target. In this case, fragments of a genetic sequence are randomly sampled, and then experiments are performed on the fragments to determine which pairs of fragments come from overlapping regions [Iris 1994]. This overlap information is then used to piece together, or assemble, the fragments into an ordered layout of the fragments that covers the original genetic sequence (see [Venter et al., 2001] pp. 1308–1319 for a complete description and an example). This problem can be solved using multiple strategies but relies on the existence of multiple data types to establish a physical scaffold onto which the assembly is placed. For example, the Celera whole genome assembler (WGA) has a five-stage process: Screener, Overlapper, Unitigger, Scaffolder, and Repeat Resolver. This is a very computer-intensive process, taking days with current state-of-the-art hardware. Every overlap computed was statistically a 1-in-10^{17} event and, therefore, not a coincidental event.

The sequence assembly problem has been extensively examined and several solutions have been applied. Venter et al. [2001] assembled the human genome using the Celera WGA as described above, while a somewhat different assembly approach was taken by Kent and colleagues for the public version of the genome [Lander et al. 2001]. As additional large genomes are solved, the assembly problem will continue to be further refined and computational efficiencies achieved. This continues to be an exciting and productive interface between computer scientists, mathematicians, algoritmists, and bioinformaticists.

The assembly problem is further compounded by the fact that all the data is inaccurate (e.g., digest lengths are off by up to 10%, electrophoretically determined sequences contain 0.5% incorrect base assignments, etc.). One approach to dealing with these problems has been developed by Phil Green, the author of the Phred [Ewing et al. 1998, Ewing and Green 1998] and Phrap programs (http://www.phrap.org). Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. Phrap is a program for assembling shotgun DNA sequence data. Key features are that it allows the use of entire reads (not just trimmed high-quality parts); uses a combination of user-supplied and internally computed data quality information to improve accuracy of assembly in the presence of repeats; constructs contig sequences as a mosaic of the highest quality parts of reads (rather than a consensus); and, provides extensive information about assembly (including quality values for contig sequences) to assist in troubleshooting. It is able to handle very large datasets and was influential in the subsequent development of more complex assemblers.

Additionally, the data is partial, not all regions of the original are represented in the sample, and the orientation of fragments is frequently unknown. Other issues involve the amount and type of information gathered to infer overlaps. More information implies more confidence in the veracity of an overlap, but some false positives will occur by chance. Consequently, it is likely that when building a scaffold from ordered clone libraries, a variety of different experiments yielding heterogeneous types of information will be performed and must be used simultaneously to detect overlaps. A key analysis problem is to accurately assess the statistical significance of overlaps under some stochastic model. Another issue involves how much to sample. Statistical and biological considerations show that for large problems with moderate overlap information, one will never achieve coverage without an impractical amount of sampling. This is the gambler's dilemma: at some point one must stop rolling the dice and move to another strategy to complete an assembly project.

While preliminary solutions to genome analysis have been of great use, many challenges remain, including:

1. The scale of the problems is such that they are computationally demanding. Better algorithms and the exploitation of parallelism are required.
2. Each assembly problem has a somewhat different combinatorial structure due to variations in the experimental methods used to infer overlaps. There is clearly a central generalized assembly problem, but each variation requires its own optimization to best lever the combinatorial structure. However, many of these projects are one-time efforts. The challenge is to build software that is both general and efficient.
3. The resulting assemblages are large, and software is needed that permits one to visualize and navigate a solution. Further engineering is required to allow investigators to manipulate solutions according to their expert knowledge, and to maintain versions of the data and a record of the work.
4. Finally, the central assembly problem involves NP-hard combinatorial problems for which heuristics work well on typical data. But as the scale and number of the problems to be solved increase, we will need ever more trustworthy solutions [Goldberg et al. 1995].

34.6.3 Sequence Analysis

Database searching has become an essential part of modern molecular biology. Database searching is our most effective means of identifying a potential biochemical function of a newly sequenced protein or gene. It is hard to imagine the number of hours of trial and error that are eliminated by the advent of this technique.

34.6.3.1 Matching a Defined Pattern

The need for speed in database searches has led to using heuristic methods. Current research topics in this area include the improvement of the sensitivity of searching, which can be severely reduced by the overall nucleotide composition of the genomes involved. Database searches are generally conducted with a global alignment algorithm that finds the "best" alignment over the entire length of both sequences. Recent

advances in our understanding of domain structure of proteins and the intron–exon organization of genes has made it very desirable to develop a fast, sensitive local alignment search algorithm to identify regions within a pair of sequences that show the highest similarity. Currently, the most promising approach to this is an implementation of one of the rigorous dynamic programming local alignment algorithms on very highly parallel hardware [Lim et al. 1993].

Several other searching techniques are widely used by experimental molecular biologists as aids for guiding and interpreting experiments. These include things as simple as finding the highly hydrophilic and hydrophobic regions in a protein sequence or regions capable of forming helices on the surface of a protein. Finding specific patterns of codon usage in a newly sequenced gene can yield insights into its expression, and finding the intron–exon junction is necessary before the correct protein sequence can be derived from a genomic DNA sequence. There are several important research topics in searching for signals. First is the need for procedures for easily and clearly specifying very arbitrary, complex patterns in a sequence. Faster algorithms for finding these patterns are needed as well. In many cases, the patterns, or biological signals, that are being sought are too complex to be readily identified by visual inspection of sequences. This is especially true if some variation is permissible in the signal pattern. Thus, another important research topic is better algorithms for selecting the most likely patterns to be associated with a signal in a sequence. For example, given several genes known to contain exons, from a single organism, how do we discover the pattern that signals the beginning and end of each exon? A variety of tools have been applied in the large genome projects [Adams et al. 2000, Lander et al. 2001, Venter et al., 2001] and further refinement is still needed.

34.6.3.2 Alignment

Sequence alignment is an important type of searching. It is, basically, the search for the most similar juxtaposition of sequences or regions within sequences. Good alignments are necessary if our inferences about the homology of genes are to be accurate. Even more crucially dependent on good alignments are methods for reconstructing phylogenies based on sequence data [Steel et al. 1994]. Finally, some problems in identifying signal patterns are appreciably simplified given well-aligned sequences. Fortunately, the state-of-the-art in pairwise sequence alignment is well-advanced. There are rigorous algorithms for both global and local alignments of pairs of sequences. These algorithms allow both flexible and sophisticated treatment of insertion/deletion gaps. One possible topic for research in this area is the context-sensitive treatment of these gaps. This would include cases where an insertion/deletion would change the reading frame of a coding region in a gene sequence or change the relative positions of amino acids known to be essential to protein function.

Multiple sequence alignment techniques are not as far advanced as the pairwise techniques. The rigorous pairwise algorithm can be, and has been, extended to multiple sequence problems. However, this approach requires computer time and memory proportional to the product of the lengths of the sequences. Recent advances have reduced this requirement by a large constant factor. However, even with this improvement, this approach soon exhausts even the fastest present-day computers. If a phylogeny of the sequences is available independent of the sequences themselves, it can be used to convert a sequence alignment problem into a series of pairwise problems. However, because a frequent reason for doing a multiple sequence alignment is to generate a phylogeny, this is not a general solution. Thus, a variety of heuristic algorithms are used for most multiple sequence alignment problems.

Several kinds of research are needed here. First, faster and more rigorous algorithms are required. Where algorithms are not completely rigorous, we need to characterize their performance so that we know how close to an optimal solution we can come. We also need to identify what sequence features might cause an algorithm to perform badly.

Defining Terms

API: Application programming interface. An API provides a wide and varied set of functions, messages, and structures that give applications access to the features and capabilities of the underlying

operating system. The API consists of a set of standard interfaces to such features as window management, graphics device interface, system services, multimedia services, and remote procedure calls.

Clone assembly: The essence of experimental physical mapping is the ordering and placement of fragmented regions of chromosomes in an overlapping contiguous stretch of DNA that covers the entire chromosomal interval. The DNA fragments are derived from the original DNA and each fragment propagated independently (cloned) to obtain working quantities. Fragment overlaps are determined by hybridization experiments with unique DNA sequences that will permit the unique identification of a chromosomal region. This is a computationally difficult problem.

Kalman filtering: A digital-image averaging procedure for the enhancement of the signal-to-noise ratio in noisy images. The Kalman filter is a recursive version of the true averaging procedure, and takes the form $y_1 = (i-1)y_i - 1/i$, in which the filter parameters are not constant but vary with the frame number i so that the latest averaged image y_n always equals $(x_1 + x_2 + \cdots + x_n)/n$. This gives a straightforward average over n frames, with a maximum signal-to-noise ratio, without having to prespecify n.

Linkage and linkage mapping: The proximity of two or more markers (e.g., genes, DNA sequences) on a chromosome; the closer the markers are, the lower the probability that they will be separated during repair or replication, and therefore, the greater the probability they will be inherited together. A map of the relative positions of genetic markers on a chromosome, determined on the basis of how often they are inherited together, is referred to as a linkage map.

Marker: An identifiable physical location on a chromosome (e.g., a specific identifiable sequence such as a restriction-enzyme cutting site, or a gene) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or a segment of DNA with no known coding function but whose inheritance can be followed with molecular techniques.

Mainchain-directed alignment: A technique for predicting protein structure from NMR data, based on aligning only mainchain residues and ignoring sidechain structures at the primary alignment stage. This technique ignores the complexity of the sidechain packing but suffers from distortions based on the effects of sidechain groups.

NOESY: Nuclear Overhauser Effect (NOE) Enhancement Spectroscopy. The NOE is a consequence of relaxation of dipole-coupled spins induced in magnetic resonance. The enhancement factor in NMR is the deviation from thermal equilibrium induced in one of a pair of dipole-coupled spins following the selective radiation of one of the pairs.

Phase problem: In x-ray diffraction structure determination, a collimated beam of monochromatic x-rays is directed through an object. The x-rays are scattered in all directions by the electrons of every atom in the object. The magnitude of the scattering is proportional to the size of the electron complement of atoms in the target. Because of the variable composition of the target, the scattering of x-rays appears continuous. However, when regular structures are radiated, and known heavy-metal ions are included, it is possible through a series of least-squares calculations to reassemble the phased waves of the diffracted x-rays and compute a clean electron density map.

Physical map: A map of the locations of identifiable regions on DNA (e.g., specific identifiable sequences or genes), *regardless of inheritance*. Distance is measured in base pairs. The physical map with the highest possible resolution is the entire nucleotide sequence of the chromosome.

Scanning confocal microscopy: A technique that — unlike conventional light microscopy, which focuses an optical image of a specimen on the image plane of a collection system — involves the physical scanning of the specimen with a diffraction-limited point of light. In some cases, this may be a one- or two-dimensional array of points. In such microscopy, the result of the interaction of the scanned light beam(s) with successive regions of the specimen is measured and recorded. With such instruments, digital signal processing may significantly enhance the signal-to-noise ratio, yielding three-dimentional images of extraordinary quality.

Sequence assembly: One of the most common methods of sequencing DNA is "shotgun" sequencing, in which a DNA strand is read as a series of random substrings of length 350 to 500. The reconstruction

of the sequence of the whole molecule from these random strings is referred to as sequence assembly. There is great complexity in this process, because the reactions to produce the sequencing substrate may be obtained from either strand. Therefore, when comparing two fragments, one must take into account that they could be derived from the same strand or from different strands; in the latter case, it is necessary to take the reverse complement of one of them before making the comparison.

Sequential alignment: A technique for evaluating the degree of secondary structure alignment by sequential evaluation of the root-mean-square deviation between backbone atoms. This involves identification of secondary structure, application of a clustering function to locate collections of such structures, and examining the more extended alignments outside of the initial regions.

Simulated annealing: This technique is a derivative of the Metropolis algorithm and applies statistical mechanics to optimization to many-body systems. In brief, the Metropolis algorithm is used to provide a simulation of a number of atoms in equilibrium at a given temperature. In each step of the algorithm, an atom is given a small random displacement, and the resulting change ΔE in the energy of the system is computed. If the $\Delta E \leq 0$, the displacement is accepted and the value is the basis for the next round. Through a series of probabilistic and cost functions, the algorithm optimizes at a given temperature. The simulated annealing procedure applies this function to a system that has been "melted," and the temperature is lowered until the system is frozen. It is essential that at each stage the system proceed long enough to reach a steady state.

SOAP: SOAP is a lightweight protocol for exchange of information in a decentralized, distributed environment. It is an XML-based protocol that consists of three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined datatypes, and a convention for representing remote procedure calls and responses. SOAP can potentially be used in combination with a variety of protocols.

Stochastic search algorithm: A form of optimization searching based on random sampling (Monte Carlo) methods where points v from a set V are chosen at random with probability $1/|V|$. The minimum values of $f(v)$ are recorded as the random sampling proceeds, and the sampling does not terminate arbitrarily, as might occur in a deterministic search. Simulated annealing is an example of a stochastic algorithm.

Transmission electron microscopy (TEM): A technique in which a suitably prepared sample is placed in a beam of electrons being controlled in an electric field. A moving electron has a wavelength that is inversely proportional to its momentum (mass times velocity). Therefore, the higher the accelerating voltage of the TEM, the smaller the wavelength and the higher the resolution. The modern TEM consists of an electron source, an imaging lens, and an image-recording system, all housed in a column under high vacuum. Electrons are emitted from a heated tungston filament held at a large negative potential and accelerated through voltages greater than 80 kV. The column is equipped with condenser electromagnetic lenses for focusing.

Web Services: A Web service is a software system identified by a URI (Uniform Resource Identifiers, a.k.a. URLs, are short strings that identify resources on the Web), whose public interfaces and bindings are defined and described using XML. Its definition can be discovered by other software systems. These systems can then interact with the Web service in a manner prescribed by its definition, using XML-based messages conveyed by Internet protocols.

References

Adams, M.H. et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195.

Bellon, P.L. and Lanzavecchia, S. 1995. A direct Fourier method (DFM) for x-ray tomographic reconstructions and the accurate simulations of sinograms. *Int. J. Bio-Med. Comput.*, 38(1):55–69.

Cheng, R.H., Kuhn, R.J., Olson, N.H., Rossmann, M.G., Choi, H.K., Smith, T.J., and Baker, T.S. 1995. Nucleocapsid and glycoprotein organization in an enveloped virus. *Cell*, 80(4):621–630.

Cinkosky, M.J. and Fickett, J.W. 1993. *SIGMA User Manual.* Los Alamos National Laboratory, Los Alamos, NM.

Desjarlais, J.R. and Berg, J.M. 1992. Redesigning the DNA-binding specificity of a zinc finger protein: a database guided approach. *Proteins: Struct. Funct. Genet.*, 12:101–104.

Dobson, C.M. 1999. Protein misfolding, evolution and disease. *Trends Bichem. Sci.*, 24:329–332.

Dyer, M., Frieze, A., and Suen, S. 1995. Ordering clone libraries in computational biology. *J. Comput. Biol.*, 2:207–218.

Eigenbrot, C., Randal, M., Presta, L., Carter, P., and Kossiakoff, A.A. 1993. X-ray structures of the antigen-binding domains from three variants of humanized anti-p185HER2 antibody 4D5 and comparison with molecular modeling. *J. Mol. Biol.*, 229:969–995.

Ewing, B., Hillier, L., Wendl, M.C., Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, 8:175–185.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8:186–194.

Fernandez, J.-J., Lawrence, A.F., Roca, J., GarcÌa, I., Ellisman, M.H., and Carazo, J-M. 2002. High-performance electron tomography of complex biological specimens. *J. Struct. Biol.*, 138:6–20.

Fink, A.L. 1998. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Dis.*, 3:R9–23.

Gilson, M.K., Straatsma, T.P., McCammon, J.A., Ripcoll, D.R., Faerman, C.H., Axelsen, P.H., Silman, I., and Sussman, J.L. 1994. Open "back door" in a molecular dynamics simulation of acetylcholinesterase. *Science* 263:1276–1278.

Goldberg, P.W., Golumbic, M.C., Kaplan, H., and Shamir, R. 1995. Four strikes against physical mapping of DNA. *J. Comput. Biol.*, 2:139–152.

Hammarström, P., Wiseman, R.L., Powers, E.T., and Kelly, J.W. 2003. Prevention of Transthyretin Amyloid Disease by changing protein misfolding energetics. *Science*, 299:713-716.

Iris, F.J.M. 1994. Optimized methods for large-scale shotgun sequencing in Alu-rich genomic regions, pp. 199–210. In M.D. Adams, C. Fields, and J.C. Venter, Eds. *Automated DNA Sequencing and Analysis,* Academic Press, London.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921.

Lim, H.A., Fickett, J.W., Cantor, C.R., and Robbins, R.J., Eds. 1993. In *Proc. 2nd Int. Conf. on Bioinformatics, Supercomputing and Complex Genome Analysis,* St. Petersburg, FL, July 1992. World Scientific, Singapore.

Luthy, R., Bowie, J.U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature (London)*, 356:83–85.

Magness, C., Xu, Y., and Green, P. 1993. *SEGMAP — a program for computing and displaying YAC-based STS-content maps*. Washington University School of Medicine, St. Louis.

Matise, T.C., Perlin, M., and Chakravarti, A. 1994. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature Genet.*, 6:384–390.

Pan, Y., Yuhasz, S.C., and Amzel, L.M. 1995. Anti-idiotypic antibodies: biological function and structural studies. *FASEB J.*, 9:43–49.

Robbins, R.J., Ed. 1994. Report on the Invitational DOE Workshop on Genome Informatics, 26–27 April 1993, Baltimore, MD. Genome Informatics I: Community Databases. *J. Comput. Biol.*, 1:173–190.

Rossman, M.G., Arnold, E., Erickson, J.W., Frankenberger, E.A., Griffith, J.P., Hecht, H.J., Johnson, J.E., Kamer, G., Luo, M., Mosser, A.G., Rueckert, R.R., Sherry, B., and Vriend, G. 1985. Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature (London),* 317:145–153.

Sali, A., Shakhnovich, E.I., and Karplus, M. 1994. How does a protein fold? *Nature (London)*, 369:248–251.

Schmieder, P., Hoch, J.C., Stern, A.S., and Wagner, G. 1995. Maximum entropy reconstruction of non-linearly sampled data. Keystone Symposium on Frontiers of NMR in Molecular Biology — IV, Keystone, Colorado, Apr. 3–9, 1995. *J. Cell. Biochem. Suppl.*, 21b, p. 76.

Steel, M.A., Szekely, L.A., and Hendy, M.D. 1994. Reconstructing trees when sequence sites evolve at variable rates. *J. Comput. Biol.*, 1:153–163.

Tramontano, A. and Lesk, A.M. 1995. Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. *Proteins*, 13:231–245.
Uberbacher, E. and Mural, R. 1991. Locating protein coding regions in human DNA sequences by a multiple sensor–neural network approach. *Proc. Natl. Acad. Sci. USA*, 88:11261–11265.
Venter, J.C. et al. 2001. The sequence of the human genome. *Science*, 291:1304–1351.
Waterston, R.H. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562.

Further Information

The Journal of Computational Biology is a regular source of computational molecular biology. Additional journals related to computational biology are in the late planning stages. The *Journal of Structural Biology* and *Proteins: Structure, Function and Genetics* is also a source of current work. For a comprehensive treatment of the mathematics of molecular biology, the reader is directed to *Introduction to Computational Biology: Maps, Sequences and Genomes,* by Michael Waterman, published by Chapman & Hall, 1995.

IV

Graphics and Visual Computing

Graphics and Visual Computing is the study and realization of complex processes for representing physical and conceptual objects visually on a computer screen. Fundamental to all graphics applications are the processes of modeling objects abstractly and rendering them on a computer screen. Also important are object identification, light, color, shading, projection, and animation. The reconstruction of scanned images and the virtual simulation of reality are also of major research interest, as is the ultimate goal of simulating human vision itself.

35

Overview of Three-Dimensional Computer Graphics

Donald H. House
Texas A&M University

35.1 Introduction

The name *three-dimensional computer graphics* has been used freely in the computer graphics community for many years now [Foley et al. 1990, Glassner 1995, Hill 1990, Rogers 1985, Watt and Watt 1992]. It is something of a misnomer, because the graphics themselves are not in any sense three-dimensional (3-D). Rather, the way that the graphics are generated is dependent upon the construction of a virtual 3-D model in the computer, which is then imaged via a virtual camera, usually implying a simulation of a real physical illumination process. The term *three-dimensional* merely emphasizes the fact that a simulation of a 3-D world underlies the image-making process and also that the images produced often display the kinds of foreshortening distortions apparent in photographs or perspective drawings of real 3-D scenes. This chapter is devoted to outlining the various aspects of the process of generating 3-D computer graphic images. It is meant to give the reader an overview, or "big picture," that can be filled in by reading Chapter 36 through Chapter 43 of the handbook, which provide more detailed information on specific aspects of the process.

35.2 Organization of a Three-Dimensional Computer Graphics System

A three-dimensional graphics system can be thought of as having three major components, each of which performs a distinct and clearly defined key role in the process of image generation. These three components are responsible for *scene specification, rendering*, and *image storage and display*. Figure 35.1 gives a schematic view of the process used in 3-D graphics, showing the role that each of these components plays. Each of these major components can itself be broken down into groups of important subcomponents.

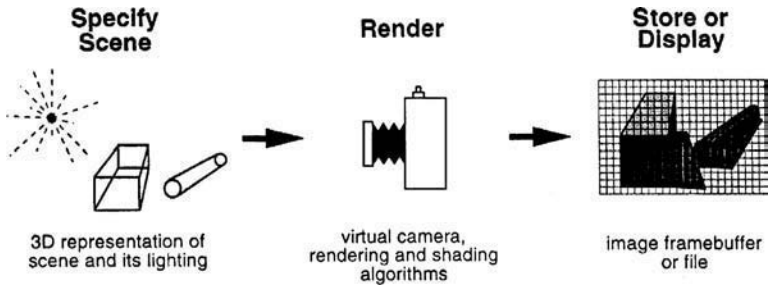


FIGURE 35.1 The 3-D graphics process.

35.2.1 Scene Specification

The *scene specification* section of a 3-D graphics system is responsible for providing an internal representation of the virtual scene that is eventually to be imaged. This requires both an interface to allow user specification and modification of the scene definition and a set of internal data structures that store and organize the scene so that it can be accessed by the user interface and the rendering system. This can be a highly interactive program, providing access to a variety of modeling tools via an interactive user interface; it may be script-driven, providing a scene description language that the user communicates in; or it can be as simple as a program that reads basic geometric information from a tightly formatted scene description file. In any case, the scene specification system will need to support some concept of a geometric coordinate system and provide some way of describing the geometry of the scene to be imaged. Scene description systems also will provide a way in which the user can specify what (virtual) materials objects are made of and how the scene is lit.

35.2.1.1 Coordinate Systems

Key to the geometric structure of a 3-D graphics system is a compact means for storing and utilizing descriptions of **local coordinate systems**. The local coordinate systems are used in the definition of the various components of a model describing the geometry and other characteristics of the scene, much as the local coordinates used on a plan are used in describing the design of a real object. For example, the coordinates on the plan for a complete airplane will necessarily be much different from the coordinates used on the plan for the airplane's wheel assembly.

Consistent with the usual representation of 3-D coordinates in mathematics and engineering, most current books, articles, and implementations of 3-D graphics systems use right-handed coordinate systems [Foley et al. 1990]. This gives a natural organization with respect to the display screen, with the x-coordinate measuring horizontal distance across the screen, the y-coordinate measuring vertical distance up the screen, and the z-coordinate providing the third spatial dimension as distance in front of the screen. However, in the early development of computer graphics, coordinate systems were often left-handed [Foley and van Dam 1982]. In screen space, the difference is that the positive z or depth coordinate is measured into the screen. Of course, for modeling, a local coordinate system can be positioned and oriented anywhere in space and is not usually aligned with the screen. Figure 35.2 shows the ordering of right-handed and left-handed coordinate systems.

A local coordinate system is usually defined in terms of a small set of intuitive geometric operations — the **affine transformations**. These are:

1. Translation — a change in the position of the origin of the local system
2. Scaling — a change in the scale of measurement in the local system
3. Rotation — a change in the orientation of the local system
4. Shear — transformation from an orthogonal coordinate system to a nonorthogonal system or vice versa via shearing deformations

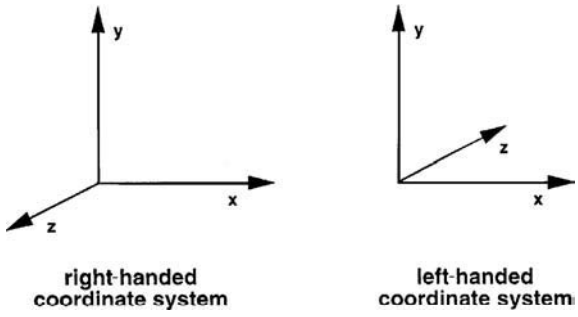


FIGURE 35.2 Right- and left-handed coordinate systems.

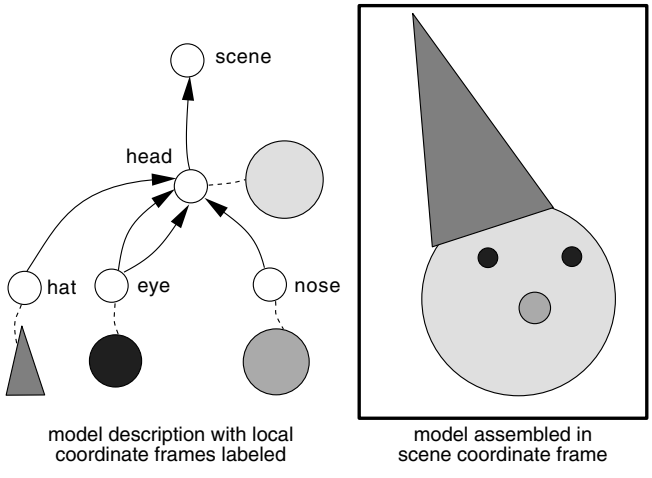


FIGURE 35.3 A clown described by a hierarchy of local coordinate frames.

All of these elements of the local-coordinate-system definition are specified with respect to the origin, scale, orientation, and shear of some external reference system, which might itself be specified relative to some other external system. In this way, local coordinates can be nested within each other, providing the possibility for models to be described in a **hierarchical** fashion. For example, the simple clown model of Figure 35.3 is described in terms of a hierarchy of coordinate frames, which allows for the design and modeling of the head, eye, nose, and hat in their own separate local coordinate frames but then places the two eyes, the nose, and the hat on the head with respect to the head's frame. Finally, the assembled head is placed and oriented in the scene with respect to a local reference frame for the scene. The reference frame at the top of such a hierarchy is usually referred to as the **global coordinate system**.

The basic geometric unit is the 3-D point, which is typically represented in a 3-D graphics system as a 3-vector and stored as a an array of three elements, representing the x, y, and z components of the point. Orientation vectors, like normals to surfaces and directions in space, are also represented by 3-vectors. Thus, the point (x, y, z) is given by the vector

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

(or in some systems by its transpose $[x \;\; y \;\; z]$).

A local-coordinate-system is usually specified by a four-dimensional (4-D) *homogeneous transformation matrix* of the form

$$M = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

which specifies a transformation from the local coordinate system to its reference coordinate system. In other words, applying the transformation M to a point specified in the local coordinate system will yield the same point specified in the reference coordinate system. Another way of thinking of the same transformation matrix M is that when applied to the reference coordinate system it aligns it with and scales it to the local coordinate system. The 4-D homogeneous form of the transformation matrix M allows the unification of translation with scaling, rotation, and shear in a single matrix representation.

The transformation implied by matrix M is actually implemented by a three-step process. Assuming that 3-D geometric points are represented as column vectors in the local coordinate system, they are transformed into the reference coordinate system by:

1. Extending the 3-D point $\mathbf{p}$ into a 4-vector $\mathbf{v}$ in homogeneous space by giving it a fourth, or w, coordinate of 1:

$$\mathbf{p} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad \Longrightarrow \quad \mathbf{v} = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

2. Premultiplying this extended vector by the matrix M yielding a transformed 4-vector $\mathbf{v}'$:

$$M\mathbf{v} = \mathbf{v}' = \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix}$$

3. Converting the resulting 4-vector $\mathbf{v}'$ into the transformed 3-D point $\mathbf{p}'$ by discarding its w-coordinate:

$$\mathbf{v}' = \begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} \quad \Longrightarrow \quad \mathbf{p}' = \begin{bmatrix} x' \\ y' \\ z' \end{bmatrix}$$

Inspection of matrix M will show that it is defined to always send the original w-coordinate to itself, thus making the third step legitimate. (In earlier computer graphics systems, it was usual for points to be represented by row vectors instead of column vectors and for step 2 to be done by *postmultiplying* the homogeneous row vector $\mathbf{v}$ by the transpose of matrix M.)

The basic transformations of translation, rotation, scaling, and shear are given by the following matrices, which assume that points are represented as column vectors in a right-handed coordinate system and that transformations will be done by premultiplication of the vector (extended to homogeneous coordinates) by the matrix. Use of left-handed coordinates instead of right-handed coordinates will affect the rotations only as indicated below. If row vectors and postmultiplication are being used to represent points and their transforms, the matrices must be transposed.

- *Translation* by Δx in the x-direction, Δy in the y-direction, and Δz in the z-direction:

$$T(\Delta x, \Delta y, \Delta z) = \begin{bmatrix} 1 & 0 & 0 & \Delta x \\ 0 & 1 & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad T(\Delta x, \Delta y, \Delta z) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x + \Delta x \\ y + \Delta y \\ z + \Delta z \end{bmatrix}$$

- *Scaling* of s_x in the x-direction, s_y in the y-direction, and s_z in the z-direction:

$$S(s_x, s_y, s_z) = \begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad S(s_x, s_y, s_z) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} s_x x \\ s_y y \\ s_z z \end{bmatrix}$$

- *Rotation* through angle θ around the x, y, or z axis, with right-handed rotation around the axis taken as a positive rotation (i.e., aligning the thumb of the right hand with the axis, the fingers grasp the axis in the direction of positive rotation; note that if left-handed coordinates are being used, the signs of the *sine* terms in R_x and R_y should be reversed, but R_z is unaffected):

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta & 0 \\ 0 & \sin\theta & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad R_x(\theta) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y\cos\theta - z\sin\theta \\ y\sin\theta + z\cos\theta \end{bmatrix}$$

$$R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\theta & 0 & \cos\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad R_y(\theta) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x\cos\theta + z\sin\theta \\ y \\ -x\sin\theta + z\cos\theta \end{bmatrix}$$

$$R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & 0 \\ \sin\theta & \cos\theta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad R_z(\theta) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x\cos\theta - y\sin\theta \\ x\sin\theta + y\cos\theta \\ z \end{bmatrix}$$

- *Shear* parallel to the (x, y) plane as a function of z, or parallel to the (y, z) plane as a function of x, or parallel to the (z, x) plane as a function of y:

$$H_{xy}(h_{xz}, h_{yz}) = \begin{bmatrix} 1 & 0 & h_{xz} & 0 \\ 0 & 1 & h_{yz} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad H_{xy}(h_{xz}, h_{yz}) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x + h_{xz}z \\ y + h_{yz}z \\ z \end{bmatrix}$$

$$H_{yz}(h_{yx}, h_{zx}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ h_{yx} & 1 & 0 & 0 \\ h_{zx} & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad H_{yz}(h_{yx}, h_{zx}) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ y + h_{yx}x \\ z + h_{zx}x \end{bmatrix}$$

$$H_{zx}(h_{xy}, h_{zy}) = \begin{bmatrix} 1 & h_{xy} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & h_{zy} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \qquad H_{zx}(h_{xy}, h_{zy}) \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x + h_{xy}y \\ y \\ z + h_{zy}y \end{bmatrix}$$

More complex coordinate transformations, involving combinations of the basic transformations, can be obtained by specifying them as a sequence of operations, each described by a 4-D transformation matrix. The product of these matrices will yield a compound transformation matrix that has the same effect as each transformation applied separately. For example, a rotation of 30° around the x axis followed by a translation to the point $(10, 20, -10)$ would be given by

$$M = T(10, 20, -10)R_x(30^\circ)$$

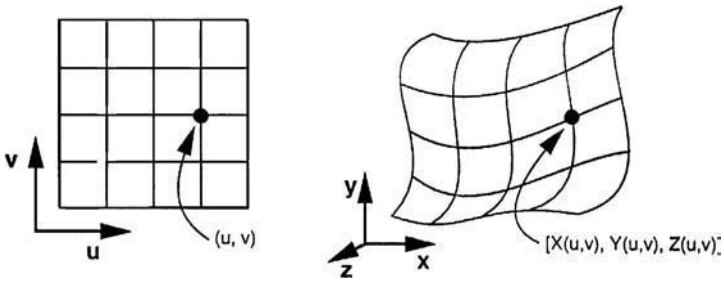


FIGURE 35.4 A biparametric surface.

35.2.1.2 Geometric Modeling

Virtually all 3-D graphics systems provide the ability to work with simple geometric primitives that can be specified as lists of 3-D points. These primitives include points, lines, and polygons. Points can be arranged together to indicate a *sampled* surface, lines to form a *wireframe* representation, and polygons to form *polyhedral* surfaces. More sophisticated modelers will provide **parametric surfaces**, which are defined via an underlying piecewise polynomial formulation [Rogers and Adams 1990, Bartels et al. 1987]. Polynomial coefficients are adjusted to give the surface a specific shape, and these coefficients are often given intuitive form by encoding them via simple geometric devices, such as control polyhedra.

A typical surface formulation is a *biparametric surface*, which describes a surface in three spatial dimensions (x, y, z) via a set of three functions of two parameters u and v:

$$x = X(u,v), \qquad y = Y(u,v), \qquad z = Z(u,v)$$

This concept is illustrated in Figure 35.4. The rectangular grid on the left of the figure, defined in the (u, v) parametric coordinate system, is mapped, via the functions X, Y, and Z, into a 3-D surface like that shown on the right. A set of points on a parametric surface can be obtained algorithmically by looping over a collection of sample points on the (u, v) plane. Simple geometric primitives and parametric surfaces are described more fully in Chapter 36.

Implicit surfaces are a common alternative to parametric surfaces. Here, surfaces are defined as the set of points satisfying a mathematical expression of the form

$$F(x, y, z) = 0$$

Thus, these surfaces are defined implicitly. Any point (x, y, z) in 3-D space can be tested to determine whether or not it is above $[F(x, y, z) > 0]$, below $[F(x, y, z) < 0]$, or on $[F(x, y, z) = 0]$ the surface. However, it is not generally easy to algorithmically generate a set of points guaranteed to be on the surface, without resorting to iterative search or relaxation techniques. Implicit formulations are especially useful for defining solids, where the implicit equation can be evaluated to determine whether a point is inside, outside, or on the surface of the solid. For example, the well-known equation for a sphere of radius r centered at the point (x_0, y_0, z_0) can be written implicitly as

$$(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2 - r^2 = 0$$

Implicit techniques can be generalized to describe surfaces or solids defined algorithmically in the form of a programmed function. This technique is useful in describing a variety of natural-looking forms [Ebert 1994]. Functional techniques are described in Chapter 37.

Some geometric data come naturally in the form of a set of scalar values distributed in a 3-D field. For example, data from medical scanners, such as MR or CT devices, consist of a set of material density values distributed on a regular lattice within a volume of space. This type of data has its own specialized set of modeling and visualization techniques that are described thoroughly in Chapter 41.

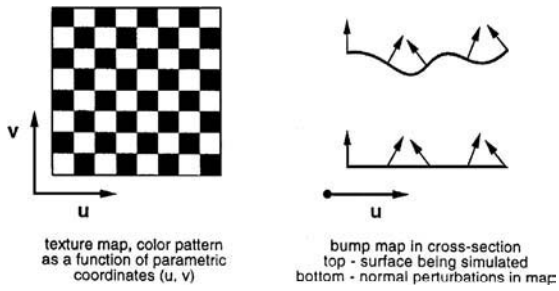


FIGURE 35.5 Texture and bump maps.

35.2.1.3 Materials

In the context of a 3-D graphics system, a **material** is an attribute of a geometric object that provides a description of how the surface of the object will appear when viewed from a particular direction under a particular illumination. In physical terms, what we need to define here is how a surface reflects (or transmits) light as a function of incident angle, reflection (or refraction angle), and wavelength. A function providing these relationships is known as the material's **bidirectional reflectance distribution function** or BRDF.

In practical terms for computer graphics applications, it is usually enough to approximate the BRDF for a material with a collection of parameters and maps. A usual material specification system will provide parameters for the specification of a material's color, diffuse reflectance factor, specular reflectance factor, specularity, transmissivity, and refraction index. These factors and their use in lighting calculations are described in detail in Chapter 38.

A material specification will also often include the capability to provide both **texture maps** and **bump maps**. A texture map provides a pattern of color that is to be applied to the surface of an object during the rendering process. These can be anything from a digital image that will be projected onto the surface, to a regular geometric pattern like a checker-board or polka-dot design. A bump map is a pattern of perturbations to the normal vector to a surface that simulates the effect that bumps would have on the appearance of the surface. Figure 35.5 illustrates texture and bump maps. Some systems also provide for **displacement maps**, which are used to locally perturb both the surface normal and the surface itself. It is usual to relate texture and other map coordinates directly to the parametric coordinates of a parametric surface. For nonparametric surfaces, specific texture coordinates must be provided by the user or by some algorithmic technique. For example, many modelers allow the user to provide texture coordinates along with 3-D geometric coordinates for each vertex in a polygonal surface.

Within the field of computer graphics, color is becoming as complex a topic as it is in physics, psychology, and art. However, from the point of view of usual practice, color is most often represented in 3-D graphics systems in one of two related color systems, the *RGB* and the *HSV* systems.

The RGB or "red–green–blue" system is the usual system used for storage or display of color. This is because this color system relates directly to the three-electron-gun RGB organization of color CRT displays. An RGB color is stored as a triple of three numbers giving the relative amount of each of the three color primaries — red, green, and blue. Because the RGB system organizes color into three *primaries*, and allows us to scale each primary independently, we can think of all of the colors that are represented by the system as being organized in the shape of a cube, as shown in Figure 35.6. We call this the RGB color cube or the RGB color space (when we add coordinate axes to measure R, G, and B levels). Note that the corners of the RGB color cube represent pure black and pure white; the three primary colors red, green, and blue; and the three secondary colors yellow, cyan, and magenta. The diagonal from the black corner to the white corner represents all of the gray levels. Other locations within the cube correspond with all of the other colors that can be displayed.

The HSV or "hue–saturation–value" color system represents colors using three measures that relate directly to how artists often think about color. It provides separate measures of *hue* (corresponding to

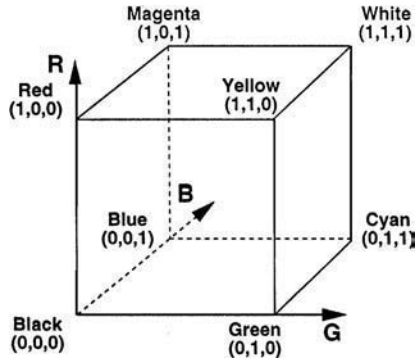


FIGURE 35.6 RGB color cube.

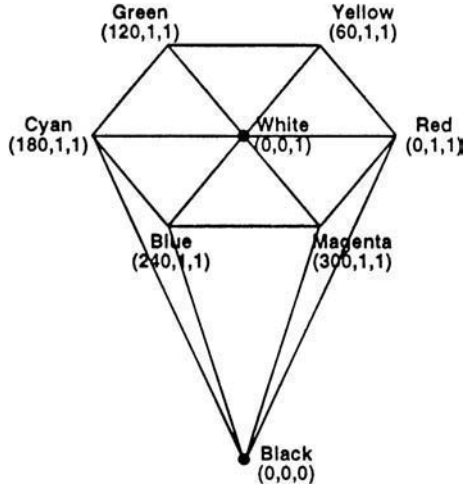


FIGURE 35.7 HSV color cone.

dominant color name), *saturation* (purity of color), and *value* (brightness on gray scale). Its structure is derived directly from the RGB system, and in fact there is a simple translation from RGB to HSV and back. If the RGB color cube of Figure 35.6 is viewed along its white–black diagonal, it presents a hexagonal silhouette. "Peeling" off layers of the face of the RGB cube visible along the white–black diagonal and projecting these cross sections onto a flat surface result in a series of smaller and smaller hexagonal cross sections. If these cross sections are then stacked up, they form a hexagonal cone. The HSV system is the cone-shaped space derived in this way, shown in Figure 35.7. Figure 35.8 shows how the coordinates of the HSV color space are organized. The hue or h coordinate is an angular measurement (usually in degrees) around the face of the cone, with red at 0°, green at 120°, and blue at 240°. The saturation or s-coordinate is measured from the center of the face of the cone out to its perimeter, with 0 at the center corresponding with gray and 1 at the perimeter corresponding with a fully saturated color of the chosen hue. The value or v-coordinate is measured from the apex of the cone to its face along the central axis, with 0 at the apex corresponding with no illumination (black) and 1 at the face corresponding with full illumination.

35.2.1.4 Lights

The purpose of lights in a 3-D graphics system is to provide the illumination source for the simulated shading calculations done by the renderer in making an image. Thus, all light sources must define a

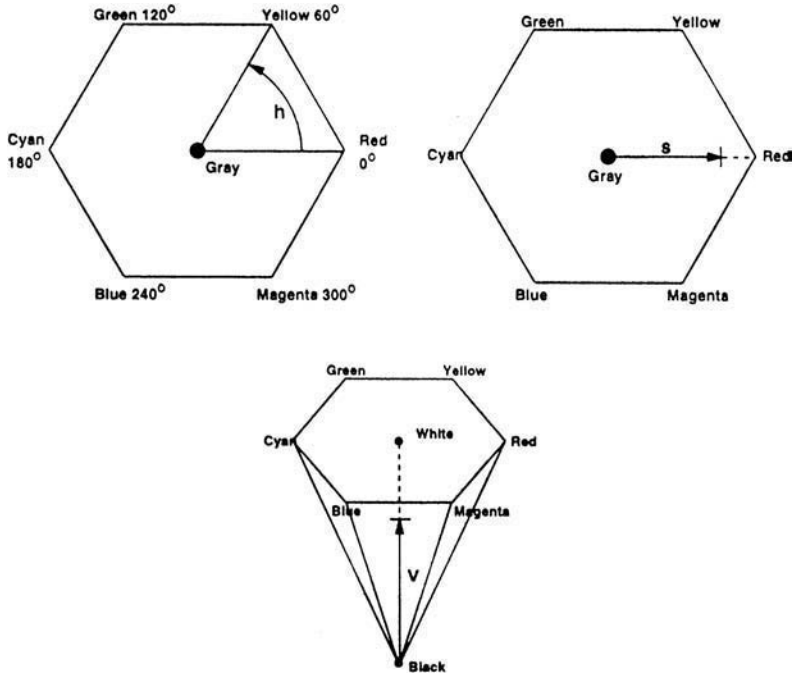


FIGURE 35.8 HSV parameterization of hue, saturation, and value.

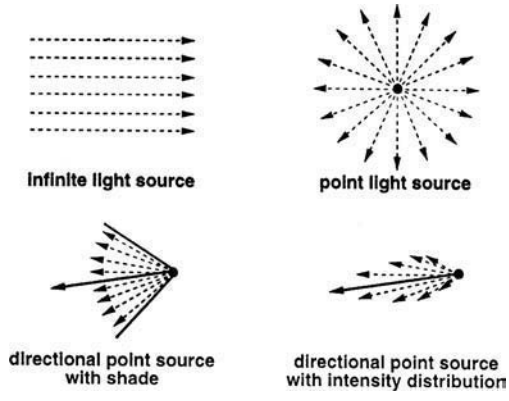


FIGURE 35.9 Infinite and point light sources and variants.

color of the illumination that they provide, usually specified in RGB or HSV coordinates. Note that this illumination color combines the notions of the intensity of the light and its chromatic attributes. Lights are arranged in a scene along with geometric objects but usually carry no geometric properties other than their position and, for directional lights, a direction of orientation. Some rendering algorithms work with light sources with other geometric properties such as shape and area, but most 3-D graphics systems work with only two types of lights — **infinite lights** and **point lights**. Figure 35.9 illustrates infinite and point light sources and some of their variants.

An infinite light is one that is so far away from the geometry being illuminated that light rays can be assumed to be parallel to each other, like the rays of light from the sun illuminating the earth. For this type of light, no position needs to be specified. All that matters geometrically is the direction of the light rays.

A point light source is assumed to have no area, with light emitted in all directions from the geometric position of the light. Simple variants of point light sources include the addition of conic or other shading devices with the light specification, so that the light shines only in a particular direction. A further variation is to have the intensity of light rays fall off gradually as a function of angular distance from the central directional axis of the light. With these variations, a point light can provide a reasonable approximation to an unshaded incandescent bulb, a shaded desk or studio lamp, a flashlight, or a spotlight.

35.2.2 Rendering

Rendering is simply the process of transforming a 3-D scene description into a two-dimensional (2-D) image. It is generally done by a simulation of the physical process that occurs in a camera when a picture is recorded on film. Making this process algorithmic, so that it can be simulated efficiently and accurately in a computer, is the essence of the rendering problem. Chapter 38, describes practical approaches to rendering in some detail, so a brief synopsis will suffice here.

Briefly, the main steps in the rendering process are:

1. Point of view — orienting the 3-D scene as if it were being viewed from a particular point in space,
2. Projection — associating points in the 3-D scene with their images on a 2-D virtual image plane or screen by projecting the 3-D scene description onto the plane,
3. Visible-surface determination — deciding which surfaces projected onto the image plane would actually be visible from the present viewpoint,
4. Sampling — fixing a set of sample points across the virtual image plane, usually corresponding in some way with the pixels that will be used to store the calculated image, and associating these sample points with visible points on the scene's 3-D geometry,
5. Shading calculation — determining, for these sample points, what color would be reflected or transmitted to the viewpoint from the geometry visible at the sample point, taking into account the scene's geometry, lighting, and materials,
6. Image construction — from the shaded samples, determining and storing colors for each pixel in the output image.

These steps are not necessarily completed in this order. Nevertheless, every renderer will have to solve each of these problems in some way.

35.2.2.1 Camera

The role of the virtual camera in a 3-D graphics system is to provide both a point of view from which to render an image and the basic parameters of the mathematical projection that will be used to form the virtual image. The camera's position and orientation are specified sometimes as part of the scene description system and sometimes elsewhere. Nevertheless, it is typical for the camera to be positioned in the global coordinate system, usually with some positioning controls that correspond to the operation of a real studio camera. Often, a camera is positioned by one set of controls and aimed by another set. Aiming is usually made easier by the option to select a **center of interest**, which is a reference point in the scene toward which the camera will orient itself.

Theoretically, cameras can have any projection characteristics, corresponding to the entire variety of lens types, either real or imagined. However, practical 3-D graphics implementations usually implement only the standard parallel or perspective projections that are common in architectural and design drafting.

A perspective projection is one in which all light rays coming from the scene converge at a common point, known as the **center of projection**. If a projection plane is interposed between the scene and the center of projection, the point at which a ray from the scene through the center of projection intersects the projection plane is the *image* of that point. Figure 35.10 shows the geometry of this projection. The parallel projections can be considered as special cases of the perspective projections, where the center of projection is infinitely far from the scene and virtual screen. Thus, rays between points on the scene and the center of projection are parallel to each other when intersecting the screen.

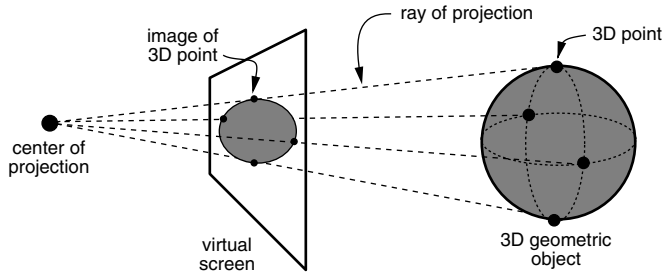


FIGURE 35.10 Geometry of a perspective projection.

In order to completely specify the perspective projection, the position of the camera (i.e., the center of projection), the direction in which the camera is aimed (i.e., its central ray of projection), the camera's up direction, the distance of the virtual screen from the center of projection along the central projection ray, and the width and height of the virtual screen must be known. This assumes that the virtual screen is centered on the central ray of projection, with its surface normal aligned with the central ray (i.e., it is perpendicular to the central ray). It is also possible to build fancier cameras, where the screen can be moved off center and oriented skewed to the central ray.

35.2.2.2 Renderer

The renderer in a 3-D graphics system is essentially the engine that drives the picture-making process. We can think of the renderer as viewing the scene through the lens of the virtual camera and constructing an image of what it sees, by first sampling points on the scene geometry and calling on the shader to calculate colors for each sample, and then combining these sampled colors into the pixels of the image. There are so many approaches to rendering, and the subject is so complex, that we will direct the interested reader to Chapter 38 for more information.

35.2.2.3 Shader

The shader is the algorithm that uses the information collected by the renderer about a point sample on the scene geometry, its material attributes, and the available lighting to calculate a color for the sample. Generally this is done by a more or less approximate physical simulation of how light is reflected toward the camera from the position on the surface at which the sample is being taken. Again, the reader is referred to Chapter 38 for more detailed information on shading and how it is done in a typical graphics system.

35.2.2.4 Image Construction

The final step in the rendering process is the construction of a digital image from the set of shaded samples across the virtual screen. This is done in any of a variety of ways, all of which are forms of low-pass filtering and resampling, providing a smooth blending and interpolation of the color samples into image pixel values [Wolberg 1990]. In practical terms, the digital image pixel grid is superimposed over the virtual screen, so that its pixels become associated with locations on the screen. Then the color of each pixel in the grid is calculated by taking a weighted average of the shaded samples in the vicinity of the pixel.

35.2.3 Storage and Display

For a 3-D graphics system to be useful, there must be a way to turn the results of calculations into tangible images that can be both viewed and stored for archiving and transmission. Thus, a 3-D graphics system is organized around a model of a digital image data structure, one or more image file formats, and a notion of the kind of display device that will be used to view images.

35.2.3.1 Image

The *pixmap* is the basic data structure for in-memory storage of digital images. A pixmap is simply a 2-D array of pixel values, with each pixel's value stored in units of one or more bits. Typical pixel sizes are 1, 8, 24, and 32 bits.

A pixmap that allocates only one bit per pixel is known as a *bitmap* and can be used to store only monochrome images (i.e., each pixel is either full on or full off). Pixmaps with 8 bits per pixel can be used to store up to 256 levels of gray for a shaded gray-tone image, or, if the image is colored, the 8 bits are usually used to index a *lookup table*, which is simply an array containing up to 256 RGB colors that are used in the image. Pixmaps of this type are limited to pictures with a palette of no more than 256 distinct colors, although these colors can be drawn from a much larger set of possible colors. The size of this set is determined by the number of bits per entry in the lookup table. This scheme is often supported by hardware as described below in the discussion of *framebuffers*.

Pixmaps with 24 bits per pixel normally allocate 8 bits, or one byte, to each of the three RGB color primaries, giving a color resolution of 256 levels per primary, or 16,777,216 distinct colors. This color resolution is well beyond the ability of the human eye to distinguish color differences, so that even color gradations in these images appear to be as smooth as they would be in a continuous-tone color image.

On a high-end graphics computer, it is not unusual to allocate more than 24 bits per pixel in a pixmap. The extra space can be used to store colors at higher than 8-bit resolution, which is often handy to avoid roundoff errors in image-processing operations. A common configuration is a 32-bit pixel, where only 8 bits are used for each color primary, and the additional 8 bits are used to store an *alpha* value. The alpha value is used in image-compositing operations as a measure of pixel opacity. For purposes of compositing images together, pixels with high alpha values are treated as if they were opaque and pixels with low alpha values are treated as if they were transparent. Other uses for extra bits in a pixmap are to store aspects of the geometric information of the original model, such as surface normal, object id, material type, 3-D position, etc. This information can be used in postprocessing of the image to do things like modify shading, or to add embellishments to the image that give a notion of the underlying form and structure of the geometry of the imaged objects.

35.2.3.2 Display Devices

The display device most frequently used in conjunction with a 3-D graphics system is the *CRT* or *cathode-ray tube*. A CRT works on exactly the same principle as a simple vacuum tube. A schematic diagram of the organization of a monochrome CRT is shown in Figure 35.11. Electrons traveling from the negatively charged cathode toward the positively charged plate are focused into a beam by focusing coils. The plate end of the CRT is a glass screen coated with phosphor. The grid control voltage adjusts the intensity of the beam and thus determines the brightness of the glowing phosphor dot where the beam hits the screen.

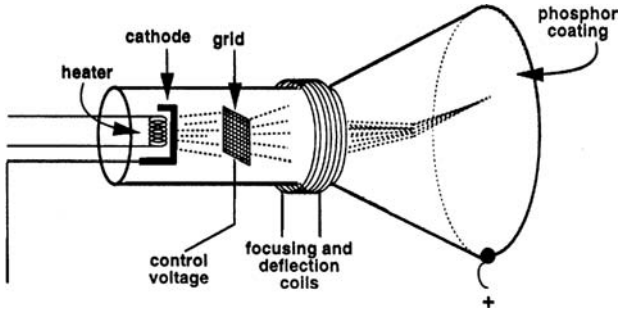


FIGURE 35.11 Schematic diagram of a CRT.

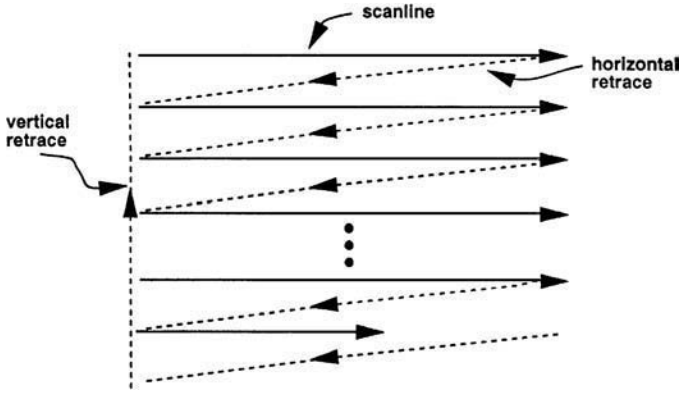


FIGURE 35.12 Raster scan pattern.

Steering or deflection coils push the beam left/right and up/down so that it can be precisely directed to any desired spot on the screen.

A color CRT works like a monochrome CRT, but the tube has three separately controllable electron beams or *guns*. The screen has dots of red-, green-, and blue-colored phosphors, and each of the three beams is calibrated to illuminate only one of the phosphor colors. Thus, even though beams of electrons have no color, we can think of the CRT as having red, green, and blue electron guns. Colors are made using the RGB system, as optical mixing of the colors of the tiny adjacent dots takes place in the eye. Typically, the colored phosphors are arranged in triangular patterns known as *triads*, and an opaque *shadow mask* is positioned between the electron guns and the phosphors to ensure that each gun excites only the phosphors of the appropriate color.

A CRT can be used to display a picture in two different ways. The electron beam can be directed to "draw" a line-drawing on the screen — much like a high-speed electronic Etch-a-Sketch. The picture is drawn over and over on the screen at very high speed, giving the illusion of a permanent image. This type of device is known as a *vector display* and was quite popular for use in computer graphics and computer-aided design up until the early 1980s. By far the most popular type of CRT-based display device today is the *raster display*. These work by scanning the electron beam across the screen in a regular pattern of *scanlines* to "paint" out a picture, as shown in Figure 35.12. The resulting pattern of scanlines is known as a **raster**. As a scanline is traced across the screen by the beam, the beam is modulated proportional to the intended brightness of the corresponding point on the picture. After a scanline is drawn, the beam is turned off and brought back to the starting point of the next scanline. As opposed to a vector display, which essentially makes a line drawing on the screen, a raster display can be used to paint out a shaded image.

The NTSC broadcast TV standard that is used throughout most of America uses 585 scanlines with 486 of these in the visible raster. The extra scanlines are used to transmit additional information, like control signals and closed-caption titling. The NTSC standard specifies a *framerate* of 30 frames per second, with each *frame* (single image) broadcast as two *interlaced fields*. The first of each pair of fields contains every even-numbered scanline, and the second contains every odd-numbered scanline. In this way, the screen is *refreshed* 60 times every second, giving the illusion of a solid flicker-free image. Actually, most of the screen is blank (or dark) most of the time. High-quality color CRTs for computer graphics greatly exceed the resolution and framerate of the NTSC standard, offering noninterlaced framerates of 60 or more frames per second with 1000 or more scanlines per frame.

35.2.3.3 Framebuffers

A *framebuffer* is the hardware interface between the pixmap data structure of a digital image and a CRT display. It is simply an array of computer memory, large enough to hold the color information for one

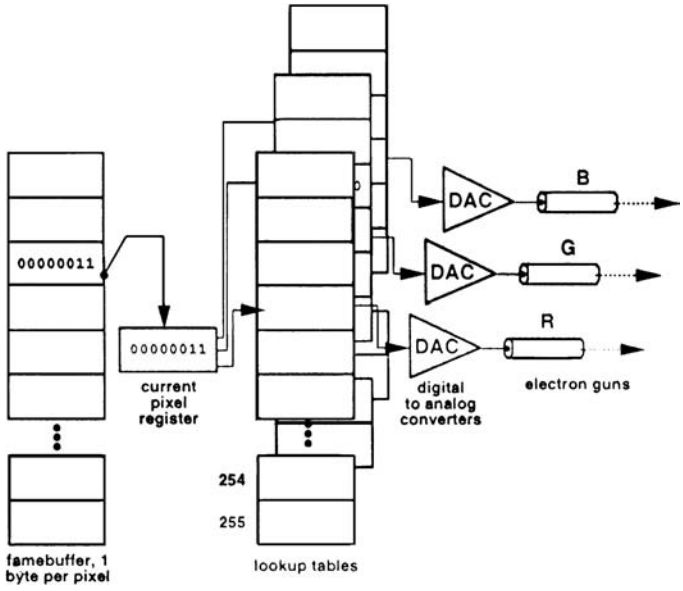


FIGURE 35.13 8-bit color framebuffer with lookup table.

or more frames (i.e., screenful) and display hardware to convert the current frame into control signals to drive a CRT. The color framebuffer schematized in Figure 35.13 holds an 8-bit per pixel color image in a pixmap. The circuitry that controls the electron gun on the CRT loops through each row of the image array, fetching each pixel value in turn and using it to index an array of 256×3 high-speed hardware registers arranged as a lookup table. The values fetched from each of the three indexed registers are converted to voltages by digital-to-analog converters (DACs) and used to control the grid voltages of the CRT's three RGB electron guns. The timing has to be such that the memory fetches and conversion to grid voltages are synchronized exactly with the trace of the beam across the corresponding screen scanline, so that the correct position on the screen is associated with the appropriate pixel from the framebuffer.

A full-color-resolution framebuffer is shown in schematic form in Figure 35.14. This type of device will have at least 24 bits per pixel (8 bits per color primary), driving three color guns, either directly or (as shown in the figure) through a separate lookup table per color primary. In this case, the lookup table is not used to increase the color resolution but instead can be used to correct nonlinearities or to obtain certain effects like overlay planes or pseudocoloring. Higher-end framebuffers may have more than 24 bits allocated per pixel, for hardware handling of such tasks as image compositing, depth buffering for hidden-surface resolution, double buffering for real-time animated display, and overlays.

35.2.3.4 Image Files

Due to the potential for using huge amounts of space, image file storage is a very important issue in computer graphics. A TV-resolution image has about $\frac{1}{3}$ million pixels — so a full-color RGB TV image will contain $3 \times \frac{1}{3} = 1$ million bytes of color information. Now, at 1800 frames (or images) per minute in a computer animation, we can expect to use up most of a 2-gigabyte disk for each minute of animation. Fortunately, we can do somewhat better than this by various file compression techniques, but disk storage space remains a crucial issue. Related to the space issue is the speed-of-access issue — that is, the bigger an image file, the longer it takes to read, write, and display.

For purposes of this overview, we will look closely at only a very simple, but very widely used image file format that has no intrinsic notion of compression. The *PPM*, or *portable pixmap*, format was devised to be an intermediate format for use in developing file-format conversion systems [Murray and vanRyper

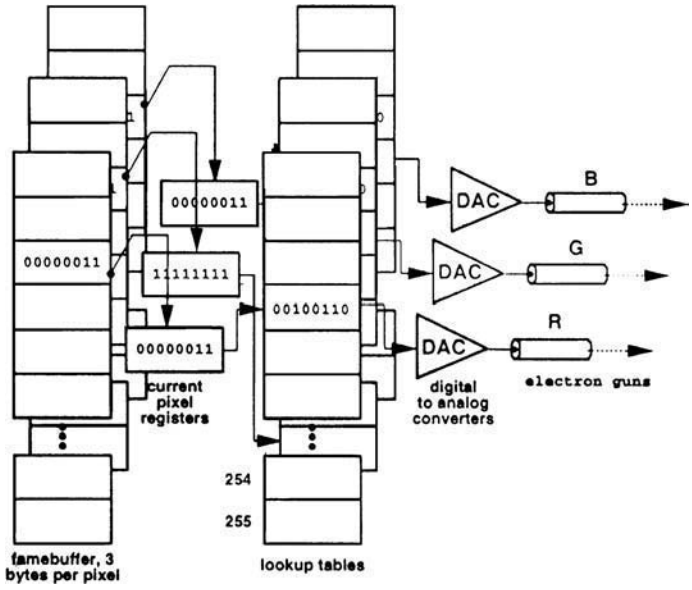


FIGURE 35.14 24-bit color framebuffer with three lookup tables.

1994]. Anyone familiar with computer graphics knows that the number of popular image file formats is immense. These include GIF, Targa, RLA, SGI, PICT, RLE, RLB, and many more. Converting images from one format to another is one of the common tasks in computer graphics, because different software packages and hardware units require different file formats. If there were N different file formats, and we wanted to be able to convert any one of these formats into any of the other formats, we would have to have $N \times (N - 1)$ conversion programs. The PPM idea is that we have one format that serves as a common source or target for all format conversions. We then write only N programs to convert all other formats into PPM and N more programs to convert PPM files into all other formats. In this way, we need write only $2 \times N$ programs to build a complete image conversion library.

The PPM format is not intended to be an archival format, so it does not need to be too storage-efficient. Although it is one of the simplest formats, PPM will nevertheless serve to illustrate some common features of image files. Most file formats are variants of the following organization. The file will typically contain some indication of the file type (which has come to be known as the format's *magic number*), a block of header or control information, and the image description data. Most (but not all) formats have a magic number, which identifies the file type. Often the magic number is not a number at all but rather a string of characters. The header block contains various descriptive information necessary to interpret the data in the image data block or can contain such archival information as creation date, image name, etc. The image data block is some form of encoding of the pixmap that describes the image. Some formats are much more complex than this, but this is the basic layout of a high percentage of formats.

In the PPM format, the magic number is either P1, P2, P3, P4, P5, or P6. P1 and P4 indicate that the image data are in a bitmap. These files are called PBM (portable bitmap) files. P2 and P5 are used to indicate gray-scale images or PGM (portable graymap) files. P3 and P6 are used to indicate full-color PPM (portable pixmap) files. The lower numbers — P1, P2, P3 — indicate that the image data are stored as ASCII characters; i.e., all numbers are stored as character strings. This has the advantage that you can read the file in a text editor. The higher numbers — P4, P5, P6 — indicate that image data are stored in a more compact binary encoding. We will look here only at P6-type files.

The header for a PPM file consists of the information shown in Figure 35.15, stored as ASCII characters in consecutive bytes in the file. In the header, all whitespace is ignored, so the program that writes the file can freely intersperse spaces and line breaks. The image width and height determine the length of a