Other Statistical Methods and Prediction Models

In modern drug discovery research, databases of drug-like properties, including partition coefficient, ionization constant, and aqueous solubility, are commonly established for a large number of compounds with diverse structures. Such databases can be utilized to develop models for predicting solubility by statistical methods. As more and more molecular descriptors are available, conventional regression methods are becoming less likely to be successful for modeling since the number of independent variables cannot be larger than the number of data points of the dependent variable for conventional regression analysis. For handling an unconventionally large number of descriptors, the following two statistical methods are particularly useful. These two methods are PLS and artificial neural network (ANN).

The method of PLS bears some relation to principal component analysis; instead of finding the hyperplanes of maximum variance, it finds a linear model describing some predicted variables in terms of other observable variables. It is used to find the fundamental relations between two matrices (X and Y), that is, a latent variable approach to modeling the covariance structures in these two spaces. A PLS model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space.

Bergstrom et al. (2004) developed in silico protocols to predict aqueous solubility of drugs. They used the solubility data of 85 compounds covering the drug-like space as identified with the ChemGPS methodology. Two-dimensional molecular descriptors for electron distribution, lipophilicity, flexibility, and size were calculated by Molconn-Z and Selma. Monte Carlo simulations in macromodel were used to obtain global minimum energy conformers, and three-dimensional descriptors of molecular surface area properties were calculated by Marea. PLS models were obtained by using training and test datasets. Both a global drug solubility model and subset-specific models (after dividing the 85 compounds into acids, bases, ampholytes, and nonproteolytes) were generated. Furthermore, the final models successfully predicted the solubility values of external test sets taken from the literature. The results showed that homologous series and subsets could be predicted with high accuracy from easily comprehensible models, whereas consensus modeling might be needed for datasets with large structural diversity.

ANN is an interconnected group of artificial neurons that use a mathematical or computational model for information processing on the basis of a connectionist approach for computation. It is an adaptive system that changes its structure on the basis of external or internal information that flows through the network, as shown in Figure 3.4.

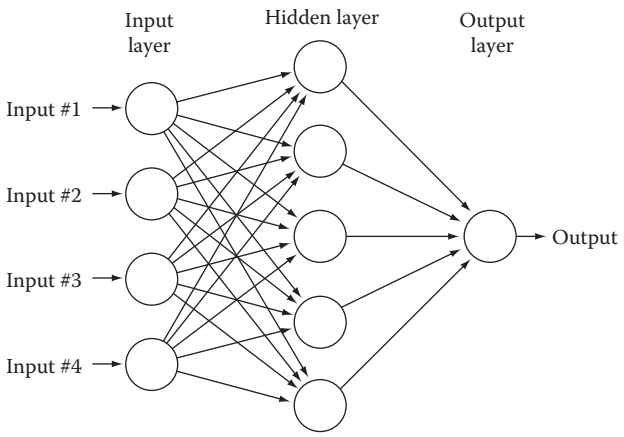

FIGURE 3.4 Schematic of the neural network structure.