Evolutionary and genetic methods in drug design

Abby L. Parrill

Many phases of rational drug design involve finding solutions to large combinatorial problems for which an exhaustive search is intractable. A simulation of the evolutionary pressure of natural selection can be incorporated into artificial intelligence algorithms to rapidly find good, if not optimal, solutions to such problems. This review describes implementations and select applications of genetic algorithms and evolutionary programming in various aspects of rational drug design. Evolutionary methods have been developed in the areas of pharmacophore elucidation, lead discovery and lead optimization, as well as in many areas of peripheral importance to rational drug design.

The evolutionary pressure of selection and the evolutionary operators crossover and mutation, which are based on ideas first described by Charles Darwin[1], can be simulated in artificial intelligence algorithms to optimize solutions to a wide variety of problems[2–4]. These evolutionary methods are often successful in finding near-optimal solutions to very complex problems.

Evolutionary algorithms

The basic steps used in evolutionary algorithms follow the pattern shown in Figure 1. An initial population of individual solutions is generated, usually by a random process. The fitness of each individual is then evaluated via a 'fitness function', which takes as input a candidate solution and returns a numeric score. Selection criteria are subsequently applied to choose individuals based on their fitness score for breeding. Finally, breeding functions are applied to produce new solutions, which replace the parent solutions. The cycle continues with the evaluation of these new solutions and continues for either a fixed number of cycles or until a particular criterion is met.

Genetic methods typically operate either in generation replacement mode (shown in Figure 1) or in steady-state mode[3]. Generation replacement involves selection of a pool of breeding individuals, whose children replace them. Hence, the entire population is turned over every generation. Steady-state mode produces one or more children for each selected parent. These children replace the least fit members of the population before the next parent(s) is(are) selected.

Some common terminology is used in reference to genetic algorithms (GAs). Each individual solution is called a chromosome by analogy to the natural model. Each chromosome consists of genes, each gene representing a parameter or feature of the solution. Different values of any gene may be referred to as alleles of that gene. Chromosomes are commonly a collection of integers or binary numbers, but can also be virtually any other type of information. Evolutionary programming (EP) primarily utilizes the term candidate solution, thus representing a phenotype rather than a genotype.

Selection functions

The basic GA cycle shown in Figure 1 has several components, which differ very little between different applications. One of these is selection of 'fit' solutions that will be allowed to breed. Common selection functions include

Abby L. Parrill, Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA. tel: +1 517 355 9715 ext. 223, fax: +1 517 353 1793, e-mail: parrill@argus.cem.msu.edu

  1359-6446/96/$15.00. PII:S1359-6446(96)10045-3

roulette-wheel selection, tournament selection and elitism. Roulette-wheel selection randomly chooses members for breeding with selection probabilities for each solution proportional to its fitness. A simple implementation of tournament selection randomly compares the fitnesses of two members of the population, retaining the most fit member for breeding. A more rigorous implementation consists of pairwise comparisons of fitness between each member of the population and one or more other randomly chosen members of the population, with the survivors being selected by the number of competitions won. Elitism is often used in conjunction with one of these methods, ensuring that the best solution in any generation is automatically carried into the next generation.

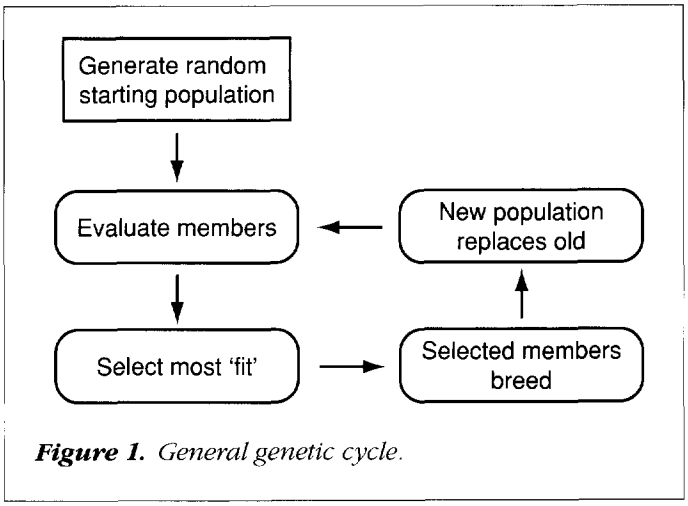

Figure 1. *General genetic cycle.*

Breeding functions

Once breeding members have been selected, breeding functions are applied. In keeping with the natural model, these algorithms are based on either sexual or asexual reproduction. Algorithms modeled after sexual reproduction (GAs) utilize crossover to combine genetic material from two parents, with mutation occurring at a low rate. Algorithms based on asexual reproduction (EP) use single parents and apply mutation at a much higher rate. The crossover function mimics the exchange of genetic material between two parent chromosomes that occurs during meiosis. The crossover function selects a random cut-point and combines the first portion of one parent with the second portion of the other to produce the first child. The first portion of the second parent is combined with the second portion of the first parent to produce the second child. This process is shown graphically on binary strings in Figure 2. The mutation function emulates the mutation of DNA, in which one nucleotide replaces the correct one, causing a change in the gene. In binary implementations of evolutionary algorithms this is done by changing a bit in the chromosome from on to off, or vice versa. Integer and floating point implementations of evolutionary algorithms involve changing a value to a different one within the allowed range, usually utilizing either a uniform or a gaussian probability distribution centered about the starting value.

Reviews covering the use of GAs in chemistry[5], molecular modeling[6], molecular recognition and design[7], and protein structure prediction[8] are available. For an application in protein design, see Schneider and Wrede[9] and the associated rebuttal[10]. This present review will concentrate on the use of evolutionary algorithms in drug design.

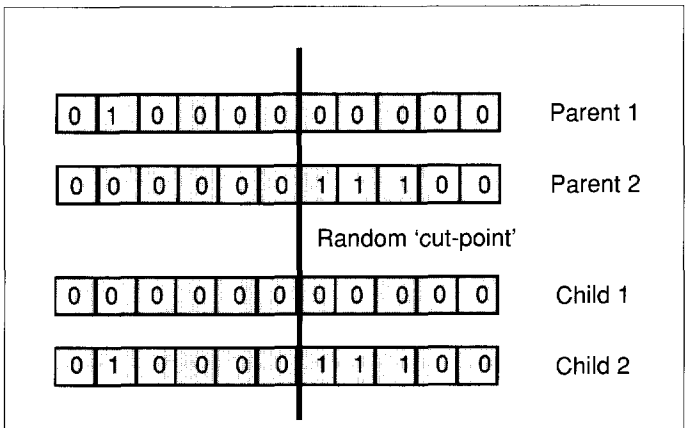

Figure 2. *Graphic demonstration of the crossover breeding function.*

Pharmacophore elucidation

An important step in the process of rational drug design is determination of the pharmacophore that elicits a particular biological response. Molecular modeling and visualization of an enzyme–ligand crystal structure provides detailed information on the pharmacophore, but is limited to available crystallized systems. QSAR studies[11] and the active analog approach[12] provide information about the pharmacophore from a series of compounds for which information on biological activity is available. A common assumption in all these studies is that all compounds in the series interact at the same biological target. Evolutionary algorithms have been applied to the problem of relating structural information to activity or property information in both a quantitative[13–17] and a qualitative manner[18–21].

Table 1. Comparison of evolutionary methods applied to the Selwood dataset for development of QSAR models

Model	Best *r* (correlation coefficient)	Corresponding *xr* (crossvalidated *r*)	Terms in fitness function
GFA	0.849	0.804	LSE: no. of terms, no. of functions
EP	0.849	0.804	RMS_ERROR: no. of terms, exponent weight
Systematic /evolution	0.909	Not provided	FIT or s_{PRESS}
GFA/NN and EP/NN	0.919	0.866	x*r*

EP, evolutionary programming; GFA, genetic function approximation; LSE, least-squares error; NN, neural network; RMS, root mean square.

Quantitative structure–activity relationships

Evolutionary algorithms have been applied to the development of QSAR models primarily as a feature selection strategy. The first application of GAs to feature selection appeared in 1992 (Ref. 14). Leardi showed that multiple linear regression (MLR) on features selected by GAs provided better models than those developed for features selected by stepwise regression. Rogers and Hopfinger utilized GAs in a similar application to select functions of one or more features in an approach they call the genetic function approximation (GFA)[16]. Available functions include linear polynomials, quadratic polynomials, splines and gaussians. Least-squares regression is used to generate the coefficients for each function in a particular chromosome. The chromosome (QSAR model) is evaluated by a lack of fit function, which is based on the least-squares error (LSE) with a modification to prevent overfitting. Tests on the Selwood[22] and other datasets showed that GFA could produce multiple, high-quality QSAR models quickly. Luke later applied EP to the feature selection problem, with coefficient calculation by least-squares regression[15]. The fitness function used in evaluating these QSAR models included the root mean square (RMS) error between predicted and measured activities, a weighting factor for the number of features in the model, and a weighting factor for the exponent used with each feature. EP was equally able to find a significant number of high-quality QSAR models for the Selwood dataset, finding some missed by the GFA method on the same datasets. It should be mentioned that this is in some part due to the stochastic nature of both algorithms.

Kubinyi utilized a combination of systematic and evolutionary methods to generate QSAR models[13]. The systematic search was used to identify all three-variable models for the Selwood dataset. Variables included in the top 25 models were selected for use in the evolutionary development of larger models. This strategy for limiting the number of variables provided considerable speed enhancement relative to a previous study[23]. Two fitness functions were used in conjunction with this method: FIT, based on the Fischer significance value (F), and s_{PRESS}, the standard deviation of predictions.

Finally, So and Karplus applied both GFA and EP as feature selection tools for the development of QSAR models using neural networks (NNs)[17]. Several fitness functions were examined, including one based on the residual error of the training set, one based on the overall error of the test set, and one based on the crossvalidated correlation coefficient (x*r*). The final two were preferred as they provide a better measure of predictivity. NNs are able to model both linear and non-linear data, and were able to determine high-quality models for the Selwood dataset used in the previously described studies.

Table 1 provides a comparison of the best QSAR models found by these methods for the Selwood dataset. The apparently superior performance of the GFA/NN and EP/NN method of So and Karplus is somewhat misleading as their method utilized x*r* as its fitness criterion, and was therefore optimizing x*r*. The other methods utilized other fitness criteria and may have found the most outstanding method by their own measures.

Active analog approach

Utilizing active analogs to elucidate the bioactive pharmacophore involves two significant steps. First, appropriate conformations and three-dimensional alignments of the various molecules must be determined. Additionally, pharmacophoric elements, which may include hydrogen bond donors, hydrogen bond acceptors, π-systems, various volume elements and electrostatic properties, must be identified.

Payne and Glen reported a method in which a GA was used to determine conformation and alignment of molecules given appropriate constraints[20]. These constraints could be derived either from pharmacophoric hypotheses or

from an X-ray crystal structure. This method, therefore, should also be applicable to the docking problem described in the next section. Subsequent work based on this method utilized a GA to determine not only conformations and alignments but also feature correlations between molecules[21]. Applications of this method to eight pharmacophore elucidation problems demonstrated that the algorithm is extremely rapid and can propose pharmacophores consistent with crystal structures for enzyme–ligand complexes and QSAR studies. One potential addition to this method the authors would welcome is the ability to utilize inactive analogs in the development of pharmacophoric models. This ability is available in the CLEW program, which utilizes a GA to learn rules relating biological activity to structural features[18]. Although CLEW does not yet carry the process through to a complete pharmacophore hypothesis, the rules generated by CLEW can be utilized to identify important features for generating molecular overlays as input to comparative molecular field analysis (CoMFA)[24] or other three-dimensional QSAR methods.

A preliminary application of two-dimensional substructure matching as a tool for pharmacophore elucidation has been reported[25]. Maximum common substructure comparison between morphine and methadone suggests groups that interact with common sites in the receptor, but, as the authors indicate, comparison of three-dimensional structures is required for a complete understanding of biological activity. Future extension of the method to three-dimensional comparisons by GA will widen the scope of the method.

Lead discovery and design

Evolutionary algorithms have been applied in the field of rational drug design to generate new lead compounds. Lead compounds can be found in existing chemical databases by fast searching or docking protocols, synthesized and isolated by combinatorial chemistry, or designed *de novo* by computational design programs. Evolutionary algorithms have been applied in each of these research areas.

Vast arrays of chemicals are compiled in databases such as the Chemical Information System[26], Beilstein[27], the National Cancer Institute Drug Information System[28] and many proprietary databases. Available information on the biological activity for these compounds is incomplete, providing impetus for the development of efficient, accurate and rapid similarity searching and computational docking protocols. Similarity searching can help find new leads with a structure or structural features similar to that of known compounds, whereas docking methods can help to identify potential drug candidates with novel structures or to optimize existing leads for biological targets of known three-dimensional structure.

Database searching

GAs have been applied to the problem of finding two-dimensional matches to a target query in chemical databases[29,30]. One implementation indicated that GAs are much more applicable to finding the maximal common subgraph isomorphism (hyperstructure matching) than to finding subgraph isomorphism (substructure matching)[29]. Fontain reported that GAs were able to find the minimum chemical distance (MCD) between isomeric molecular ensembles, particularly when molecular connectivity information was incorporated to produce a knowledge-augmented GA (Ref. 30). The MCD was applied to elucidation of the reaction mechanism, but could also be applicable to similarity searches if broadened to comparison of non-isomeric molecular ensembles. A related type of database searching is utilized for selection of maximally dissimilar molecules for biological testing in primary screens to identify types of molecules that should be further investigated as leads. A comparison of GA selection of dissimilar molecules with selection by a newly described centroid algorithm has indicated that the GA method is both less effective and less efficient than the newer algorithm[31].

GAs have also been applied to the problem of comparing three-dimensional structures, both to determine optimal alignments of molecular electrostatic potential fields in rigid searches[32] and in flexible searching for a pharmacophoric pattern[33]. The latter was performed in three stages. The first phase involves database screening by smoothed, bounded distance matrices. The second phase is a geometric search of remaining compounds to eliminate additional structures. The final phase is a conformational search to confirm that molecules can adopt stable conformations that fulfill the pharmacophoric constraints. GAs were one of five conformational search methods evaluated and gave results second only to the directed-tweak methods, but they do have greater flexibility for non-distance query constraints such as grid- or surface-based electrostatic properties. Although Clark and coworkers first applied conformational search GA methodology to pharmacophoric pattern matching, other groups have also implemented GA-based conformational search methods[34–38].

Docking

Another method that can identify new lead compounds involves docking ligands into a biological target, a subject that has been reviewed recently[39]. Target structures can be obtained from X-ray crystallographic studies, NMR experiments, homology modeling or computerized generation of receptor models. Walters and Hinds have applied a GA to the generation of atom-based receptor models[19]. Their method, GERM, assumes the active conformations and alignment of ligands, and generates surrounding atoms that emulate potential receptor configurations. Genetically evolved receptor models determined for a series of 22 sweet-tasting compounds provided calculated binding energies that correlated with sweetness. Such models might be used as the biological target in docking studies.

Given an appropriate three-dimensional biological target structure, an ideal solution to the docking problem would be to quickly determine the conformation and orientation of a flexible ligand with respect to a flexible host that gives the most favorable Gibb's free energy for the binding process (ΔG_B) without user definition of the binding site or important interactions. The ΔG_B depends not only on the enthalpic interactions between the ligand and host, but also on conformational enthalpy changes and entropy changes resulting from changes in solvation and conformational degrees of freedom[40,41]. The difficulties inherent in this rigorous treatment of the docking problem cause most researchers to work with various simplifications.

Initial efforts toward applying GAs to docking problems focused on optimizing the orientation of a rigid ligand with respect to a rigid host[42,43]. The fitness of individual solutions was based on the interaction energy between the host and the ligand. Even with the simplification of disregarding molecular conformations and entropic effects, structures of aromatic clusters[42] and the actinomycin D complex with deoxyguanosine[43] were determined from random initial relative orientations. The program developed for these studies has been made available for a nominal fee[44].

Recently, several groups have published reports describing the application of evolutionary algorithms to the docking problem[45–52]. Each of these newer methods incorporates ligand flexibility, and one[50] allows some flexibility of binding site side-chains in the host molecule. The EPDOCK program combines an EP search of orientational and conformational degrees of freedom of the ligand within the rigid binding site[51,52]. This program determines fitness of docked configurations using an intermolecular steric and hydrogen bonding potential with consideration of ligand internal energy based on a torsional potential and constant penalty for intramolecular close contacts. This simplified estimate of ΔG_B allowed the rapid and accurate determination of the docked configurations of ligands in dihydrofolate reductase and HIV-1 protease. The intermolecular steric term in the fitness function is scaled throughout a run to allow initial ligand orientations to overlap with host atoms and to evolve into the binding cavity.

A second approach that allows ligands to evolve into the binding site is the 'Grow' method[48,49]. This method first finds favorable conformations and orientations of a small fragment of the ligand, growing the remainder of the ligand over several generations. Validation of the method included docking ligands into thermolysin, carboxypeptidase A and dihydrofolate reductase, in comparison with the crystal structures of the complexes and previous docking studies.

Clark and Ajay have developed DIVALI (docking with evolutionary algorithms), which incorporates a novel masking operator to direct the search for binding modes[45]. Masking restricts subpopulations by freezing one bit in the binary representation of the individuals in the population. This was used with multiple subpopulations to ensure a thorough search as well as to restrict substructures in the ligand to particular regions of the receptor. Results for docking to the periplasmic binding protein, carboxypeptidase A, dihydrofolate reductase and ribonuclease A indicate that the method is capable of finding reasonable binding configurations with appropriately defined masking.

A fourth application of genetic methods in docking is the extension of the well-known DOCK program[46,53]. DOCK uses GA methods to dock flexible ligands into a rigid receptor after characterizing and identifying putative binding sites using a surface sphere cluster method[54]. The fitness of docked configurations is based on an intermolecular function considering van der Waals and electrostatic interactions, with ligand internal energy represented solely by a penalty for atomic contacts within the sum of the van der Waals radii. This method was evaluated with good results for docking ligands to dihydrofolate reductase, thymidylate synthase and HIV-1 protease. Applicability to database screening was also considered; however, a negative control for such an application was not provided. Effective screening requires that the best configurations of inactive, or non-binding, compounds have fitness values lower than those for good configurations of active, or binding, compounds. This point is particularly important for those methods that seek to decide whether or not a compound will bind (binding pre-

Table 2. Comparison of evolutionary methods applied to the docking problem

Method	Flexible ligand	Flexible receptor	Structure prediction	Binding prediction	Binding site prediction[a]	Fitness function terms
GAME	No	No	Yes	No	Yes	Intermolecular energy
EPDOCK	Yes	No	Yes	No	No	Intermolecular steric/hydrogen bonding
Judson *et al.*	Yes	No	Yes	No	No	Initial 'bump' check CHARMM energy
Clark and Ajay	Yes	No	Yes	No	No	Amber inter- and intramolecular energies
DOCK	Yes	No	Yes	Yes[b]	Yes	Intermolecular van der Waals/electrostatics Ligand 'bump' check
Meadows and Hajduk	Yes	No	Yes	No	No	Intermolecular restraints Repulsive van der Waals
Jones *et al.*	Yes	Partial	Yes	No	No	Intermolecular hydrogen bonding Intermolecular van der Waals Intramolecular van der Waals (ligand)

[a]Refers to ability of method to find binding site without user input other than protein and ligand structures.
[b]Evaluation focused on algorithm speed as a measure of applicability for database searching.

diction), in addition to determining the binding orientation for ligands known to bind (structure prediction).

Another application of GAs in docking is that of Meadows and Hajduk, who developed a protocol for docking with experimental distance restraints[47]. Docked configuration fitnesses included a van der Waals term and a distance term. This protocol was applied to optimization of ensembles of molecules with averaging of interatomic distances to allow the simultaneous determination of multiple binding sites. Although the example utilized false restraints to dock an ensemble of two ligands to one known and one false site on the biological target, it demonstrated the applicability of the method to systems with multiple ligand binding sites.

Yet another method utilizes a GA to search conformational and configurational space of the ligand while allowing terminal hydrogen bonding groups in the active site of the host to rotate in order to optimize hydrogen bonding[50]. This method uses a GA to optimize hydrogen bonding interactions between the enzyme and ligand with a fitness function, including both van der Waals and hydrogen bonding. Results obtained from docking ligands to dihydrofolate reductase, L-arabinose-binding protein, and sialidase effectively reproduced experimentally observed close contacts between enzymes and ligands. Table 2 provides a comparison of the scope of these docking methods.

Optimization of primary biological screens

A novel variation using GAs to aid the lead discovery process has been developed at Sterling Winthrop[55]. This project utilizes a GA to select a subset of rhinovirus strains that reflect broad-spectrum rhinoviral inhibitor activity for use in a primary screen during drug development. The project represents a significant collaboration between experimental and computational chemists to greatly improve the consistency of a primary antiviral screen.

De novo *design*

In addition to finding lead compounds by the previously described methods, new lead compounds can be developed by experiments in *de novo* design. Glen and Payne have applied a GA to the design of structures based on a wide variety of user-defined constraints[56]. Their method can utilize scalar constraints such as molecular weight, surface constraints such as electrostatic potential, and grid constraints such as hydrophobic/hydrophilic character of nearby atoms. Constraints selected for various design experiments provide the basis for the fitness function. The most significant difference between this GA implementation and standard implementations is the wide variety of mutation operators that incorporate chemical knowledge into the normally blind genetic optimization process. These

mutation operators include translation, molecular rotation, bond rotation, adding/removing a double bond, adding/removing fragments, adding/breaking a ring, insertion of a methylene and changing atom types. Three *de novo* design applications – design of molecules of a specific weight and shape, design of ribose replacements, and design of dihydrofolate reductase inhibitors – demonstrated the broad scope of the method.

Lead optimization

Combinatorial chemistry

One approach for the optimization (as well as discovery) of lead compounds is combinatorial chemistry, involving the synthesis and screening of large libraries of compounds to determine which exhibit biological activities of interest. Two groups have applied GAs to focus synthetic efforts in combinatorial chemistry[57,58]. Both groups showed that fairly small subsets of a combinatorial library, based either on peptides[57] or on a four-component Ugi reaction[58], can serve as a population of potential lead compounds. Biological evaluation of this population provides fitness values for a GA, which then proposes the next generation (subset of the library) for synthesis and evaluation. This method results in biological activities that improve through each generation while requiring less synthesis and testing than standard combinatorial methods.

Another application of GAs in combinatorial chemistry is the selection of fragments for assembly into the combinatorial library[59]. Sheridan and Kearsley applied a GA to the optimization of tripeptoids constructed from a wide variety of primary and secondary amines. Each tripeptoid generated during a GA run was evaluated on the basis of its similarity to a target tripeptoid or a trend vector based on activity data. Tripeptoids with high fitnesses after 25 generations were analyzed for recurring fragments, which represent the subset of primary and secondary amines which should be used during library synthesis. This method is applicable for the development of combinatorial libraries for use in specific biological screens, rather than for non-targeted screening.

De novo *design leads*

A second approach for the optimization of lead compounds is the optimization of the results of the *de novo* design program, PRO_LIGAND (Ref. 60). PRO_LIGAND generates new leads by assembling fragments from substructure libraries. These leads include high- and low-quality molecules. The low-quality molecules produced, however, may contain important structural fragments outweighed by steric interactions between other fragments and the biological target. The GA uses these new leads (both high and low quality) as the initial population for the optimization process. Results for the design of distamycin and methotrexate mimics indicate that the GA optimization step not only increases the average fitness of the solutions, but also finds better solutions than the design step.

Summary and future perspectives

Computational methods that mimic the evolutionary pressure of natural selection are applicable to a wide variety of optimization problems. Such algorithms have been used to aid the processes of pharmacophore elucidation, lead generation and lead optimization. Genetic methods are able to select important features for QSAR models, determine potential bioactive conformations and orientations for pharmacophore elucidation and similarity calculations, flexibly dock ligands into biological target structures, design new ligands *de novo*, and optimize lead compounds through cycles of combinatorial synthesis.

Genetic methods are limited in two major respects. First, they are stochastic in nature and have a finite probability of missing the optimal solution while finding near-optimal solutions. The development of more efficient non-stochastic algorithms could allow the optimal solution to be found within a reasonable time period, thus reducing the need to use the less consistent stochastic methods. Second, they require a fitness function that adequately ranks potential solutions. The greatest future improvements of GAs and EP in drug design will likely be in the ability to implement fitness functions that better reflect molecular properties as computing capabilities grow. Additionally, integrated packages incorporating evolutionary methods should become commercially available and user-friendly. These latter developments should make the use of genetic methods available to more researchers.

ACKNOWLEDGEMENT

My thanks to Dan Gehlhaar for his comments on this manuscript.

REFERENCES

1 Darwin, C. (1973) *The Origin of Species*, Dent Gordon

2 Fogel, L.J., Owens, A.J. and Walsh, M.J. (1966) *Artificial Intelligence through Simulated Evolution*, Wiley

3 Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, p. 412, Addison-Wesley

4 Holland, J.H. (1992) *Adaptation in Natural and Artificial Systems*, MIT Press
5 Hibbert, D.B. (1993) *Chemometr. Intell. Lab. Syst.* 19, 277–293
6 Brodmeier, T. and Pretsch, E. (1994) *J. Comput. Chem.* 15, 588–595
7 Willett, P. (1995) *Trends Biotechnol.* 13, 516–521
8 Böhm, G. (1996) *Biophys. Chem.* 59, 1–32
9 Schneider, G. and Wrede, P. (1994) *Biophys. J.* 66, 335–344
10 Darius, F. and Rojas, R. (1994) *Biophys. J.* 67, 2120–2122
11 Hansch, C. (1969) *Acc. Chem. Res.* 2, 232–239
12 Marshall, G.R. *et al.* (1979) in *Computer-Assisted Drug Design* (Olson, E.C. and Christoffersen, R.E., eds), pp. 205–226, American Chemical Society
13 Kubinyi, H. (1994) *Quant. Struct.-Act. Relatsh.* 13, 393–401
14 Leardi, R., Boggia, R. and Terrile, M. (1992) *J. Chemomet.* 6, 267–281
15 Luke, B.T. (1994) *J. Chem. Inf. Comput. Sci.* 34, 1279–1287
16 Rogers, D. and Hopfinger, A.J. (1994) *J. Chem. Inf. Comput. Sci.* 34, 854–866
17 So, S. and Karplus, M. (1996) *J. Med. Chem.* 39, 1521–1530
18 Parrill, A.L., PhD Thesis, University of Arizona, 1996
19 Walters, D.E. and Hinds, R.M. (1994) *J. Med. Chem.* 37, 2527–2536
20 Payne, A.W.R. and Glen, R.C. (1993) *J. Mol. Graph.* 11, 74–91
21 Jones, G., Willett, P. and Glen, R.C. (1995) *J. Comput.-Aided Mol. Design* 9, 532–549
22 Selwood, D.L. *et al.* (1990) *J. Med. Chem.* 33, 136–142
23 Kubinyi, H. (1994) *Quant. Struct.-Act. Relatsh.* 13, 285–294
24 Cramer, R.D.I., Patterson, D.E. and Bunce, J.D. (1988) *J. Am. Chem. Soc.* 110, 5959–5967
25 Wagener, M. and Gasteiger, J. (1994) *Angew. Chem., Int. Ed. Engl.* 33, 1189–1192
26 Heller, S.R. (1985) *J. Chem. Inf. Comput. Sci.* 25, 224–231
27 Hicks, M.G. (1992) *Anal. Chim. Acta* 265, 291–300
28 Milne, G.W.A. and Miller, J.A. (1986) *J. Chem. Inf. Comput. Sci.* 26, 154–159
29 Brown, R.D. *et al.* (1994) *J. Chem. Inf. Comput. Sci.* 34, 63–70
30 Fontain, E. (1992) *J. Chem. Inf. Comput. Sci.* 32, 748–752
31 Holliday, J.D., Ranade, S.S. and Willett, P. (1995) *Quant. Struct.-Act. Relatsh.* 14, 501–506
32 Wild, D.J. and Willett, P. (1996) *J. Chem. Inf. Comput. Sci.* 36, 159–167
33 Clark, D.E. *et al.* (1994) *J. Chem. Inf. Comput. Sci.* 34, 197–206
34 Meza, J.C. *et al.* (1996) *J. Comput. Chem.* 17, 1142–1151
35 Walters, W.P., PhD Thesis, University of Arizona, 1995
36 Legrand, S. and Merz, K. (1992) *FASEB J.* 6, A132
37 Judson, R.S. *et al.* (1993) *J. Comput. Chem.* 14, 1407–1414
38 Blommers, M.J.J. *et al.* (1992) *Biopolymers* 32, 45–52
39 Jones, G. and Willett, P. (1995) *Curr. Opin. Biotechnol.* 6, 652–656
40 Ajay and Murcko, M.A. (1995) *J. Med. Chem.* 38, 4953–4967
41 Balbes, L.M., Mascarella, S.W. and Boyd, D.B. (1994) in *Reviews in Computational Chemistry* (Lipkowitz, K.B. and Boyd, D.B., eds), pp. 337–379, VCH
42 Xiao, Y.L. and Williams, D.E. (1993) *Chem. Phys. Lett.* 215, 17–24
43 Xiao, Y.L. and Williams, D.E. (1994) *J. Phys. Chem.* 98, 7191–7200
44 Xiao, Y.L. and Williams, D.E. (1994) *Comput. Chem.* 18, 199–201
45 Clark, K.P. and Ajay (1995) *J. Comput. Chem.* 16, 1210–1226
46 Oshiro, C.M., Kuntz, I.D. and Dixon, J.S. (1995) *J. Comput.-Aided Mol. Design* 9, 113–130
47 Meadows, R.P. and Hajduk, P.J. (1995) *J. Biomol. NMR* 5, 41–47
48 Judson, R.S., Jaeger, E.P. and Treasurywala, A.M. (1994) *J. Mol. Struct.* 308, 191–206
49 Judson, R.S. *et al.* (1995) *J. Comput. Chem.* 16, 1405–1419
50 Jones, G., Willett, P. and Glen, R.C. (1995) *J. Mol. Biol.* 245, 43–53
51 Gehlhaar, D.K. *et al.* (1995) *Chem. Biol.* 2, 317–324
52 Gehlhaar, D.K. *et al.* (1995) *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, MIT Press
53 Dixon, J.S. (1992) in *Trends in QSAR and Molecular Modelling 92* (Wermuth, C.G., ed.), pp. 412–413, ESCOM
54 Kuntz, I.D. *et al.* (1982) *J. Mol. Biol.* 161, 269–288
55 Jaeger, E.P. *et al.* (1995) in *Computer-Aided Molecular Design* (Reynolds, C.H., Holloway, K.M. and Cox, H.K., eds), pp. 139–155, American Chemical Society
56 Glen, R.C. and Payne, A.W.R. (1995) *J. Comput.-Aided Mol. Design* 9, 181–202
57 Singh, J. *et al.* (1996) *J. Am. Chem. Soc.* 118, 1669–1676
58 Weber, L. *et al.* (1995) *Angew. Chem., Int. Ed. Engl.* 34, 2280–2282
59 Sheridan, R.P. and Kearsley, S.K. (1995) *J. Chem. Inf. Comput. Sci.* 35, 310–320
60 Westhead, D.R. *et al.* (1995) *J. Comput.-Aided Mol. Design* 9, 139–148

In short...

Incyte Pharmaceuticals, a company specializing in genomic database products, recently announced two new collaborations. A collaborative research agreement has been signed with the Structural and Genetic Information Laboratory at **Centre National de la Recherche Scientifique** (CNRS) in Marseille, France, which is headed by Dr J-M. Claverie. Incyte will fund research programs to develop new algorithms for identifying coding regions and methods to identify novel gene families.

The second collaboration is with **SCRIPTGEN**, based in Medford (MA, USA), which concentrates on the development of drugs to control gene expression, using proprietary high-throughput screening technology. SCRIPTGEN and Incyte will start a joint research program in the field of bacterial functional genomics.

Isis Pharmaceuticals has received a US patent (no. 5,563,255) for ISIS 5132/CGP69846A, an antisense inhibitor of C-raf kinase. The compound is currently in Phase I clinical trials as a treatment for a broad range of **solid tumours**, with results anticipated at the end of the year. The project is part of the antisense R&D collaboration between Isis and Ciba established earlier this year.