J Comput Aided Mol Des (2012) 26:73–75
DOI 10.1007/s10822-011-9503-4

PERSPECTIVE

Darwinian Docking

Irwin D. Kuntz

Received: 13 November 2011 / Accepted: 29 November 2011 / Published online: 6 December 2011

Abstract The Darwinian model of evolution is an optimization strategy that can be adapted to docking. It differs from the common use of genetic algorithms, primarily in its acceptance of diverse solutions over finding "global" optima. A related problem is selecting compounds using multiple criteria. I discuss these ideas and present the outlines of a protocol for selecting "hits" and "leads" in drug discovery.

Keywords Evolution · Optimization · Docking · Multivariate · Quasispecies

Darwinian evolution offers interesting insights into improved docking strategies.

This brief communication will review the characteristics of Darwinian evolution, often paraphrased as "survival of the fittest". I suggest a method of obtaining more diverse results in docking and discuss the general problem of multi-dimensional optimization in a docking context.

The evolutionary process suggested by Darwin [1] and Wallace [6] includes three essential steps: some form of selection or fitness testing, a reproductive step, and a method or methods of introducing change, typically through mutational or recombination events. Fitness refers to the continual testing of individuals according to a "fitness" or "objective" function. Individuals that pass such tests have opportunities to reproduce and thus spread nominally successful "genes" into later generations. Mutation can be generalized as a multifaceted process by which the information stored in the current population is altered through a variety of mechanisms and transmitted to future generations via inheritance during reproduction.

The whole procedure can be thought of as an optimization strategy. It has some appealing characteristics compared with more familiar optimization methods. It emphasizes local rather than global maxima and often yields a wider range of acceptable solutions than conventional strategies (including most implementations of genetic algorithms). It is a non-gradient technique that can handle a wide range of objective functions and can overcome barriers to explore a neighborhood. It is suitable for optimizing multiple components simultaneously.

The primary reasons that simulation protocols of this broad type are resistant to premature convergence lies both in the nature of the fitness functions and the nature of the selection tests. The fitness functions can contain components from many variables. The selection tests, often of a stochastic nature, provide the better ranking individuals with a higher probability of passing the tests and continuing into the next generation. Thus, there is no pressure to find the single "best" individual (to the exclusion of all others), although the "genes" of such an individual are quite likely to be maintained in the population.

Of course, we cannot expect that such an approach would outperform specialized optimization algorithms either in efficiency or in finding "global" optima. Nor is this approach directly equivalent to the "genetic" algorithms that have been used for docking: both the objective functions and the mutational strategies are different. Further, the explicit goals differ—here we do NOT seek convergence to the globally optimal compound. Global optima are a worthy goal if we had assurance that such global solutions to our nominal fitness functions would be

I. D. Kuntz (✉)
University of California, San Francisco,
San Francisco, CA, USA
e-mail: kuntz@cgl.ucsf.edu

successful drugs, drug leads, or even ligands of predictable affinity. Unfortunately, none of the existing scoring schemes can guarantee such success, and it seems unproductive to exhaustively optimize functions that are not successful. The challenge, as we all realize, is identifying drug candidates or drug leads is a multifaceted process with many variables. We need a better way of handling multi-dimensional optimization.

Computational simulations of evolution received a major impetus with the formulation of the quasi-species theory by Eigen and Schuster [2] decades after the origins of mathematical population genetics was formulated by Wright [8]. Both these theoretical approaches and much current biology indicate the importance of "biodiversity"—the spread of phenotypes (i.e. characteristics) in a population subject to multiple cycles of selection, reproduction, and mutation.

I consider two scenarios in this paper. The first is a docking scheme in which time and/or computational resources are limiting and it is not feasible to dock the entire database of compounds. Some subset of the data must be chosen. How is this task to be accomplished and still explore most of the interesting compounds? An extension of this scenario is the incorporation of computer synthesis of compounds or conformational adaptation not found in the original database.

A second scenario assumes that the full set of compounds of interest can be examined with conventional docking. Then our focus becomes how to incorporate other variables into the fitness function along with the docking scores.

Scenario I can be tackled using the evolutionary protocol as an analogy. The full database is initially clustered. A subset of cluster heads is chosen. Fitness is tested based on dock scoring and other criteria (see below). A subset of the best scoring compounds is chosen and expanded through reproduction. Then a "mutation" process is invoked that replaces some of the set with "similar" compounds: either other members of the clusters or the products of some simple (computer) synthesis procedure. Any new compounds are rescored and the cycle continues until a steady state of fitness is achieved, the time available has been used or the whole space has been explored. Given the assumption about resource limitations, it is important that the number of compounds explored at any one time be manageable. Selection can be adjusted to allow any desired fraction of compounds to continue into the next round. Monitoring the population of each candidate provides a measure of its relative merits without the need for convergence to a single compound.

In Scenario II, all the scoring and evaluation can be carried out in an a priori fashion. Because the full set of compounds has been examined there is no need for the "mutation" and "reproduction" steps. The entire focus is on the fitness tests.

The fitness tests are critical to either scenario. In the first case, they determine which compounds move forward in the evolutionary cycle. In the second case, where all the compounds can be tested, they offer a way to explore constraints beyond putative binding affinities.

The core question is how to optimize multiple uncorrelated variables. We have all faced the problem of rank ordering a list of potential hits or potential drug candidates given data on binding, cell transport, metabolism, blood levels, etc. Two common strategies are "cutoffs" and weighting schemes. Cutoffs are typically applied sequentially. We take the top X candidates from DOCK and subject them to further tests. The difficulties with this protocol include the lack of justification for a particular threshold or set of thresholds and the "AND" logic which insists that successful candidates must pass all tests rather than asking which candidates are the best in the set.

Weighting schemes are used as a routine matter in both academic and industrial settings. But we must recognize that they have neither fundamental generality nor theoretical underpinnings. The essential difficulty is that weighting schemes cannot accurately convert multi-dimensional data into one-dimensional projections. Malcolm Gladwell [4] wrote an illuminating discussion of this issue in non-mathematical terms in a recent New Yorker article on the US News and World Reports method of rank ordering colleges.

I have a simple proposal for how to set up multivariable fitness tests. Prepare a separate ordered list of the compounds for each variable. For each list devise a function that represents your current best guess for linking the value of the variable with the probability of *over-all* success in the project—e.g. given a blood level of X, what is the chance of having a commercial oral drug. Then use the conventional statistical formulas for combining independent probabilities. One also will need a "threshold" for final acceptance. I strongly encourage using a stochastic or "fuzzy" threshold in the absence of firm knowledge of what the lower limit should be. These calculations can be used in an evolutionary cycle, as noted above, to allow candidates to reproduce and continue. They can be used simply to select the best "over-all" scoring in a set of compounds where the over-all score contains components from each important variable. Lastly, this type of calculation can help evaluate existing or prospective libraries.

This proposal has positive features: it combines current best models for predicting success in a format that allows tests and refinements of each contribution. It promotes a diverse range of solutions by allowing stellar performance in one category to overcome mediocre rankings in other areas. It also allows "generalists" that score moderately

well in all areas to compete successfully. It can be put into a number of statistical frameworks, including Bayesian analysis [3, 5, 7].

There are also familiar shortcomings: the quality of the selected compounds cannot be better than the quality of the scoring functions or the probability functions. Further, the typical variables used in drug development are often correlated, and this correlation should be taken into account for a proper statistical treatment.

I offer these scenarios in the spirit of this special issue to explore new ideas. It is clearly a first approximation with the Darwinian model acting merely as a guide. Implementation and integration into existing docking schemes is obviously a significant undertaking, with many practical decisions to be made and tested. The treatment offered here is extremely sketchy and non-rigorous, but hopefully, it provides incentive to explore alternatives to the rank order methods we currently use.

Acknowledgments I am pleased to acknowledge helpful conversations with Brian Shoichet, John Irwin and their research group at UCSF.

References

1. Darwin C (1859) On the origin of species by means of natural selection. John Murray, London
2. Eigen M, McCaskill J, Schuster P (1989) The molecular quasi-species. Adv Chem Phys 75:149–153
3. Eriksson L, Antti H, Gottfries J, Holmes E, Johansson E, Lindgren F, Long I, Lundstedt T, Trygg J, Wold S (2004) Using chemometrics for navigating in the large data sets of genomics, proteomics and metabonomics. Anal Bioanal Chem 380:419–429
4. Gladwell M (2011) The order of things: what college rankings really tell us. The New Yorker Magazine, 14 Feb 2011
5. Ståhle L, Wold S (1988) Multivariate data analysis and experimental design in biomedical research. Prog Med Chem 25:291–338
6. Wallace AR, Darwin C (1858) On the tendency of species to form varieties. Linnean Society of London
7. Wold S (1991) Chemometrics, why, what, and where to next? J Pharm Biomed Anal 9:589–596
8. Wright S (1932) The roles of mutation, inbreading, crossbreeding and selection in evolution. In: Proceedings of the 6th international congress of genetics, vol 1, pp 356–366