Pattern Recognition and Distributed Computing in Drug Design

W. Graham Richards

Department of Chemistry, University of Oxford, Central Chemistry Laboratory,
South Parks Road, Oxford OX1 3QH, UK

E-Mail: graham.richards@chem.ox.ac.uk

Received: 30th April 2002 / Published: 15th May 2003

Abstract

Computational methods developed in the area of medical imaging can be adapted to find ligand binding sites on proteins. Once the binding site is specified, libraries of real or virtual molecules may be screened to seek out compounds which have very strong affinity. Massively distributed computing enables huge numbers of molecules to be screened.

These approaches will be illustrated by reference to a search for inhibitors of the binding of anthrax lethal factor to the protection antigen. With the site identified some 3.5 billion molecules were tested in 24 days using the power of 1.4 million personal computers running a screen saver. Over 300,000 hits were revealed with approximately 12,000 looking particularly promising.

Introduction

The origin of complexity in drug design is obvious: the sheer number of potential molecular structures with drug-like properties. The completion of the first stage of the human genome project revealed, somewhat surprisingly, that there are only a few tens of thousands of genes. This limits the number of protein targets which might be the starting point for drug design to at most a few hundred thousand. There are no such limitations on the molecular structures of drug candidates. Even if we demand a molecular weight range of between say 150 and 800 together with water-soluble compounds and synthesizability, only imagination limits the possibilities. There are billions of possibilities.

Two developments, one scientific and the other technological, have emerged in mitigation: pattern recognition techniques and massively distributed or grid computing.

PATTERN RECOGNITION

Although rarely invoked in the chemical literature, pattern recognition is a major topic in the field of engineering. Enormous effort has been expended in areas such as computer vision; medical imaging and perhaps above all in military applications. In a series of recent papers we have adapted some of the techniques to aspects of drug discovery.

The first problem which we have attempted to overcome using pattern recognition is that of the alignment of molecules which is the key prerequisite to inferring the nature of a binding site when we start without knowledge of the protein target structure at the atomic level. (1,2) Although there are a plethora of alignment techniques, almost all fail if one tries to superimpose one structurally optimally on top of a second structure when those two molecules are of very different sizes. Simple methods, which start by matching centres of mass, will overlay by placing the smaller molecule in the middle of the larger one. A technique from computer vision (2) which employs local structure analysis evades this trap. A sample example is shown in Figure 1.

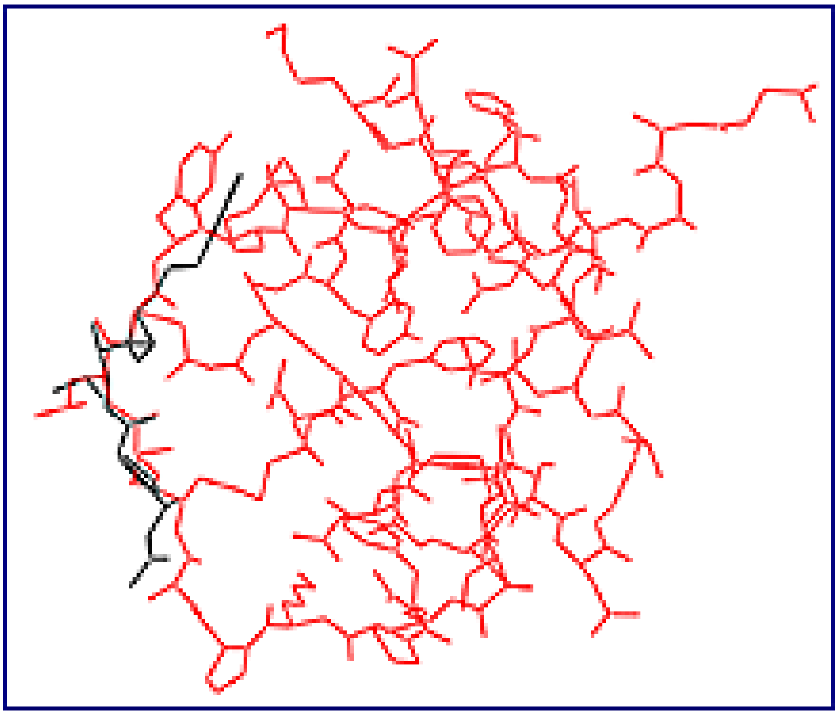

Figure 1. The computed alignment between DFKi (in black) and TOMI (in red) .

Here the larger molecule, a natural product turkey ovomucoid inhibitor is matched with the small molecule mimic DFKi which is a standard list introduced by Masek et al. (3). The small molecule is predicted to match exactly in the position where the surface active peptide fragment is situated and the matching only takes seconds on a personal computer.

More important still has been an adaptation of the multiscale approach from medical imaging (4) to find binding sites on proteins of known structure but unknown target binding site. As in medical imaging one does not want too much detail too soon or else one is overwhelmed by complexity. We need to do something very quickly and crudely but in an algorithmic stepwise manner more progressively to greater detail as we home in on the answer we are seeking. Techniques of this nature are for example used in the automatic scanning of mammograms. At the molecular level we can, in principle, find the binding site of a ligand on a protein by computing interaction energies between ligand and protein for all possible relative positions and orientations of the partners. This is the classic docking problem. Most available software essentially cheats by starting with the small molecule in roughly the correct position, perhaps already in a binding groove on the protein surface. Pattern recognition enables one to avoid this bias and still dock the pair in seconds by a hierarchical approach, illustrated in Figure 2.

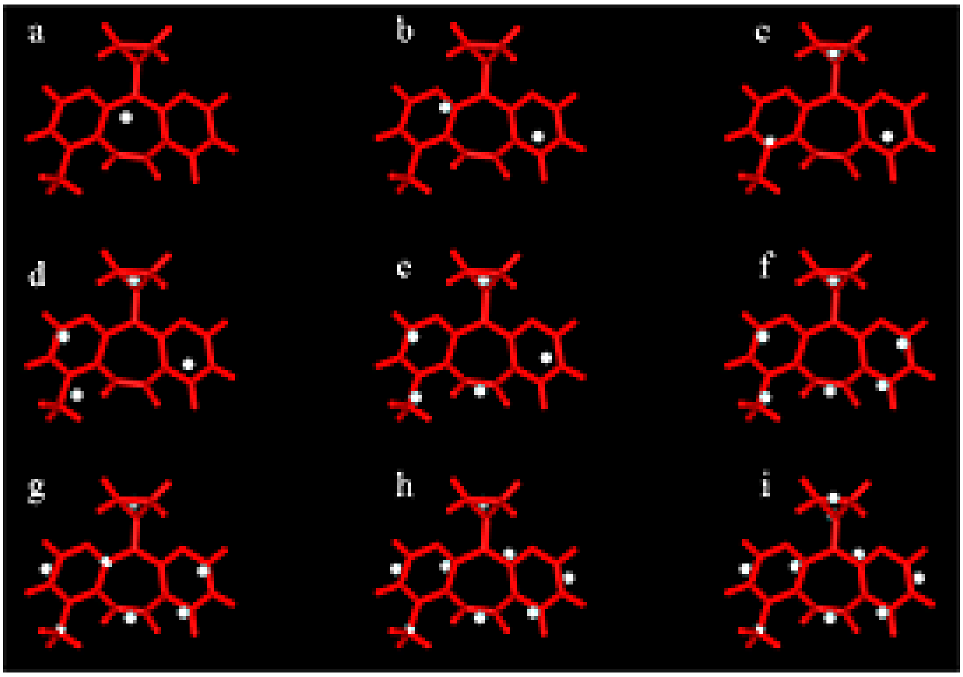

Figure 2. The hierarchy of models of nevirapine.

Firstly the small ligand is treated as a single point at the centre of mass with all the parameters for a molecular mechanics interaction energy collapsed and averaged on to that point. If we place limits on plausible interaction energy then the very rapid calculation of interaction between a point molecule and the protein will immediately restrict those parts of space where binding is not possible: not too far away to interact nor too close so as to have atom clashes. We then step up to a two-point model using the so-called k-means algorithm. To move from a one-cluster to a two-cluster or two-point model, we first find the atom furthest from the initial cluster and make this the temporary centre of the second point. All atoms closer to this centre than the original one are assigned to the second cluster and then the positions are iterated until self-consistent where each cluster centre is positional at the mean position of the atoms which belong to that cluster. This is then repeated in a stepwise fashion yielding one, two, three, four-centre models, but the interaction energy calculation is more restricted in space as the model grows in complexity. Our experience is that usually four steps are sufficient to locate the binding site to an accuracy comparable with the resolution of the crystal structure. This may then of course be optimized by employing more rigorous energy optimization procedures.

Once one has both a protein structure and a binding site, the essential feature of drug design is to try as many small molecules as possible in that binding site to see if they fit, preferably allowing flexibility in the protein and conformational freedom in the small molecule. One should then ideally compute relative binding free energies. One rapid manner of achieving these aims, albeit crudely, is to match patterns of pharmacophores. From the protein binding site structure one defines a set of binding points of obvious types: hydrogen bond donors and acceptors; lipophilic groups; changes; aromatic rings. All have positions in space. One then seeks complimentary binding contributors on the ligand for all possible shapes, allowing some leeway in distance constraints to permit a little macromolecular flexibility. When complimentary centres are found, and four such matches would seem to be ideal as this would incorporate stereochemistry in the ligand, a crude binding free energy may be computed.

The crude binding free energy will most easily be derived from a scoring function which adds up the number of specific interactions with a standard energy contribution for each (e.g. each hydrogen bond contributing 3.3 kcal mol^{-1}). The effect of the loss of conformational entropy on binding may be estimated by counting the number of rotatable bonds which are frozen and multiplying this by a factor. In addition a simple calculation using a molecular mechanics formula will give an idea of the van-der-Waals interaction energy and ligand torsional contribution. Thus matching pharmacophore patterns and then estimating energy contributions provides a measure of drug potential, although the utility will depend on the database of small molecules tried.

SMALL MOLECULE DATABASE

In our much publicised cancer drug project (www.chem.ox.ac.uk) a total of 3.5 billion small molecules have been screened as inhibitors of proteins, twelve of which are implicated as suitable targets for anti-cancer agents and also one as a site for blockers of anthrax which is discussed below.

To achieve such a large database, Davies (5) reviewed 1.4 million molecules which can be found in catalogues, and over one billion which have been made in published chemical libraries. These were then filtered to ensure that all had drug-like properties in terms of molecular weight and solubility (judged by having nitrogen or oxygen providing at least 20% of surface area). This filtering left 35 million molecules. For each of these 100 *de novo* structures were created by substitution of groups, such as $-CH_3$ for $-OH$ etc.

Although this is probably the biggest database of molecules ever screened, it must be emphasised that these lists have not been scrutinised for ease of synthesis, and one could envisage creating an even bigger and more useful database where the knowledge of organic and medicinal chemists could be incorporated. It is tempting to set up a website which is totally open where chemists the world over just contribute structures which they believe are synthesizable but with no limits on imagination. In that way one would hope to create a database with as wide a coverage of chemical space as is conceivable.

DISTRIBUTED COMPUTING

Even using fast pattern recognition techniques the screening of 3.5 billion molecules against protein targets is an enormous computational task. We have managed to achieve this using the methods of distributed computing in collaboration with United Devices Inc. (www.ud.com). The code to perform the calculations described above has been incorporated into a screensaver in a project started in April 2001. In a period of twelve months over 1.5 million personal computers have joined the project from over 200 countries, and contributed more than 100 thousand years of CPU time. It is currently (April 2002) possible to screen the database of 3.5 billion molecules in three weeks. (6)

The project is managed by a central server which dispatches work units of a number of small molecular structures and receives and validates the results. Part of the screensaver communicates with the server while the rest performs the calculations described above.

An Example: The Anthrax Project

The events following September 11, 2001 highlighted the dangers posed by the use of anthrax as a weapon of bioterror and accelerated scientific publication of relevant data. The essential molecular aspects of the action of anthrax are summarised in Figure 3.

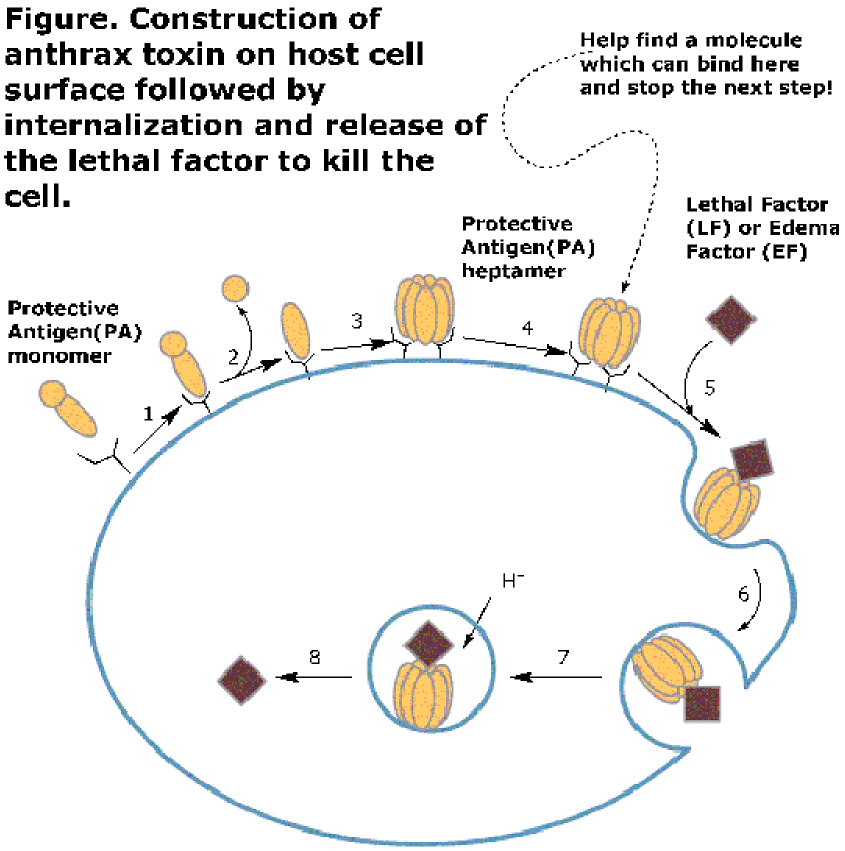
Figure. Construction of anthrax toxin on host cell surface followed by internalization and release of the lethal factor to kill the cell.
Help find a molecule which can bind here and stop the next step!
Lethal Factor (LF) or Edema Factor (EF)
Protective Antigen(PA) heptamer
Protective Antigen(PA) monomer
1
2
3
4
5
6
7
8
H⁻

Figure 3. A representation of the mode of action of the anthrax toxin.

The toxin is comprised of three proteins. The so-called protective antigen (PA) forms a heptamer which facilitates the entry into the cell of the lethal factor (LF) and the edema factor (EF): individually the proteins are non-toxic. Work by Mourez et al (7) showed that a non-natural peptide where bound to a flexible polymeric backbone inhibits the binding of PA to LF/ EF and protected rats against anthrax. Peptides which are active contained the short sequence YWWL which is not present in the anthrax proteins, suggesting that this hydrophobic peptide plays a role in binding to the PA heptamer so as to preclude the essential binding of LF/EF. One is thus in the situation described above where one has a huge protein of known structure and information about a significant ligand but no knowledge of its binding site.

Application of the pattern recognition software (8) highlights the binding site, which when once located proves to be at reasonably obvious and convincing position as illustrated in Figure 4, at the junction between proteins where LF/EF might be expected to bind but in precise atomic detail with very obvious hydrogen bonds as in Figure 5.

Armed with this information it was possible, using the distributed computing and screensaver, to try all the 3.5 billion molecules of the database. Using the then 1.4 million PCs the task was completed in 24 days, yielding some 300,000 crude hits of which about 12,000 look promising enough to merit further more refined scrutiny.

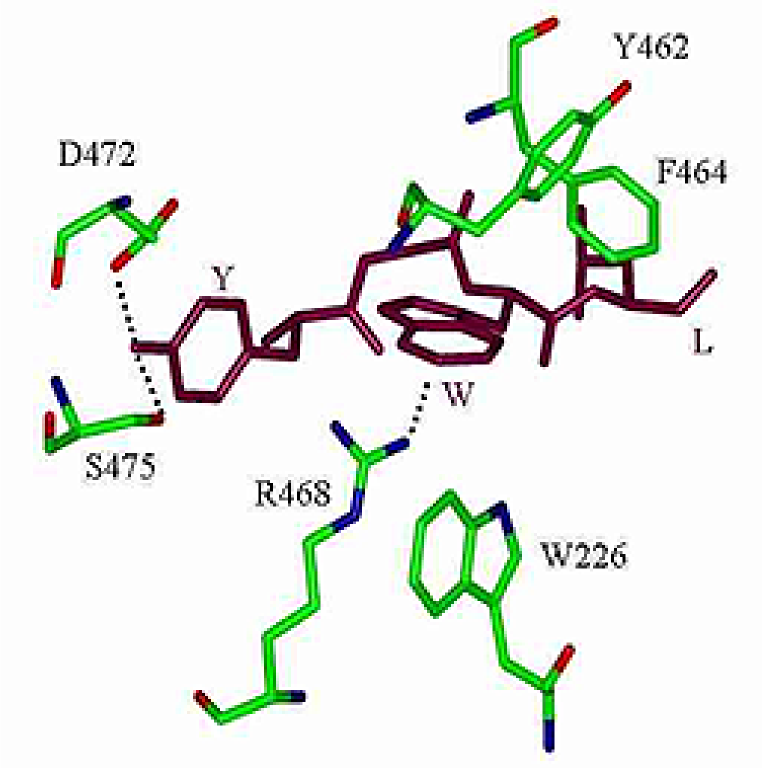

Figure 4. The heptamer of the anthrax protection antigen with the predicted bound tetrapeptide in blue.

Conclusion

The complexity of dealing with billions of potential drugs is eminently tractable if we combine the twin approaches of pattern recognition and distributed computing. Pattern recognition permits the specification of binding sites on proteins which is clearly going to become more and more important as structures flow from structural genomics. Pattern matching of pharmacophores for first-pass screening of massive databases is an ideal way to use the almost limitless power of distributed computing.

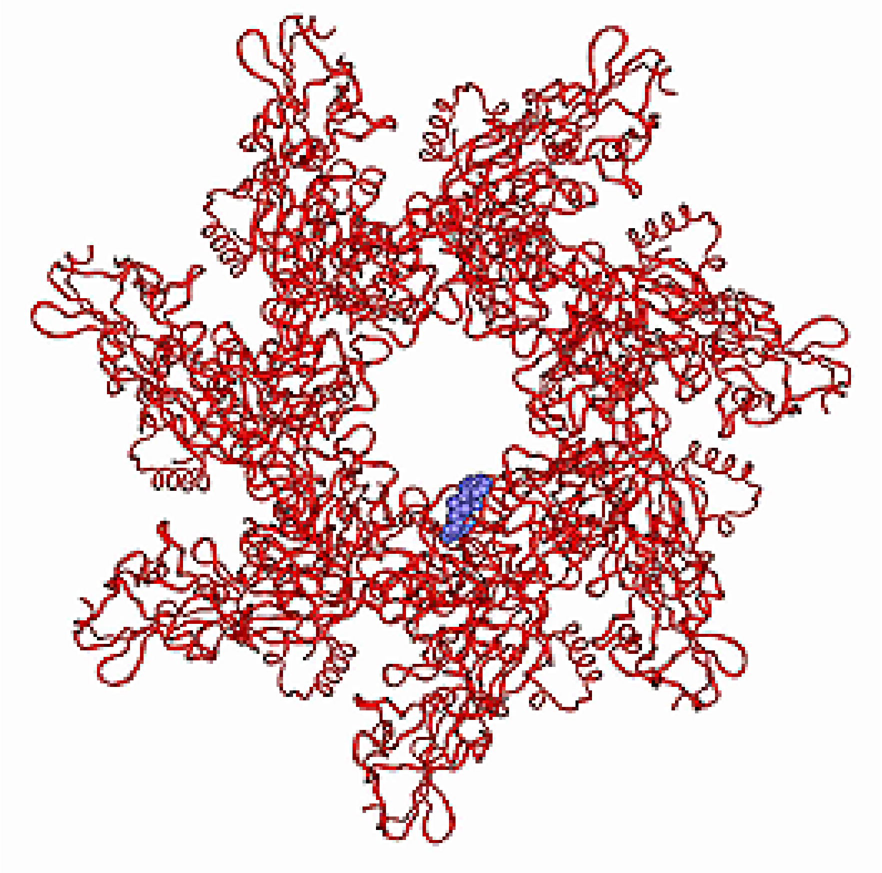

Figure 5. A detailed view of the tetrapeptide binding site of the anthrax toxin.

In the future much work of this nature is likely to be done in-house within the walls and indeed fine walls of pharmaceutical houses, but it would be a great pity if the open use of the web as employed in our cancer project were not to continue. Not only does this provide a huge, essentially free resource, the involvement of the general public as a means of increasing public understanding and indeed participation in science should not be underestimated.

Acknowledgements

This work has been supported in part by the National Foundation for Cancer Research and the ScreenSaver Projects received sponsorship from Intel Inc. and from Microsoft Corporation.

References and Notes

[1] Robinson, D. D., Lyne, P. D., Richards, W. G. (1999). Alignment of 3D-structures by the method of 2D-projections. *J. Chem. Inf. Comput. Sci*. **39**:594.

[2] Robinson, D. D., Lyne, P. D., Richards, W. G. (2000). Partial molecular alignment via local structure analysis. *J. Chem. Inf. Comput. Sci*. **40**:503.

[3] Masek, B. B., Marchant, A., Mathew, J. B. (1993). Molecular skins: a new concept for quantitative shape matching. *Proteins: Struct. Funct. Genet*. **17**:193.

[4] Glick, M., Robinson, D. D., Grant, G. H., Richards, W. G. (2002). Identification of ligand binding sites on proteins using a multiscale approach. *J. Am. Chem. Soc*. **124**:2337.

[5] Davies, E. K. submitted for publication.

[6] Davies, E. K., Glick, M., Harrison, K. N., Richards, W. G. Pattern recognition and massively distributed computing. *J. Comp. Chem*. in press.

[7] Mourez, M., Kane, R. S., Mogridge, J., Metallo, S., Deschatelets, P., Sellman, B. R., Whitesides, G. M., Collier, R.J. (2001). Designing a polyvalent inhibitor of anthrax. *Nature Biotechnol*. **19**:958.

[8] Glick, M., Grant, G. H., Richards, W. G. (2002). Pinpointing anthrax inhibitors. *Nature Biotechnol*. **20**:118