The potential of Internet computing for drug discovery

E. Keith Davies and W. Graham Richards

Large-scale, high precision drug discovery calculations, such as predicting protein folding or small-molecule protein inhibitors, have frustrated computational chemists because the supercomputers currently available are insufficiently powerful. The increasing power of PCs offers an alternative approach by harnessing 'idle time' from corporate and home computers that are connected to the Internet or an intranet. However, although the approach has the potential of offering hundreds or thousands of years of computer time per elapsed day, the architecture constraints require computational chemists to choose their methods and applications with care. Some algorithms, such as those for molecular simulations, are generally not appropriate, whereas virtual screening of molecules for protein inhibition works well.

***E. Keith Davies and W. Graham Richards**
Dept of Chemistry
Oxford University
South Parks Road
Oxford
UK OX1 3QT
tel: +44 1865 275 906
fax: +44 1865 275 905
*e-mail: keith.davies@chemistry.oxford.ac.uk

▼ The sequencing of the genome is virtually complete. The answer to the question 'what next?' is now clear to most as proteomics. The tens of thousands of genes give rise to hundreds of thousands of proteins whose individual structures will probably be determined at an accelerating rate over the next decade. But what comes after that? How is the knowledge of many, many protein structures to be exploited? The answer must be by developing small-molecule inhibitors of the functions of these proteins. However, although the genome has tens of thousands of genes and the proteome hundreds of thousands of proteins, the number of small molecules runs into billions, even if when restricted to those that have drug-like properties; this is termed the 'chemizome'.

Existing software [1–6] is used extensively within the pharmaceutical industry to screen libraries of molecular structures by docking potential inhibitors of enzymes or ligands into known protein binding sites. These virtual screening (VS) approaches are a subset of the many data mining methods [7] that are commonly used to select molecules for testing from the hundreds of thousands of small molecules in historic collections or virtual combinatorial libraries. Distributed computing offers the possibility of overcoming the limitations inherent in even the largest supercomputers, with the potential to improve the quality of hits by eliminating false positives and increasing the variety of novel leads by searching very much larger virtual collections of molecules.

New projects use screen-saver technology, such as the well-known SETI@home (Search for Extraterrestrial Intelligence at home) study, which seeks out evidence for extraterrestrial intelligence by distributing the analysis of incoming radio signals over individual personal computers (PCs) linked to the Internet (http://setiathome.ssl.berkeley.edu). Over 3.5 million participate in SETI@home, with the accumulated time of approximately one million years growing by around 1000 years per day. The CAN-DDO cancer screening project [8] (http://www.chem.ox.ac.uk) uses screen-saver time donated by individuals, with 300,000 participants joining the scheme in the first three weeks, increasing to over one million in six months. The scientific aspects of the search are coordinated from the NFCR Centre for Computational Design directed by Graham Richards of the Department of Chemistry of Oxford University (Oxford, UK), and the THINK software (E.K. Davies and C.J. Davies, unpublished) is used on massively distributed computing provided by United Devices (Austin, TX, USA; http://www.ud.com) and sponsored by Intel (Chandler, AZ, USA; http://www.intel.com/cure). Other smaller-scale projects, with up to 50,000 participants,

1359-6446/02/$ – see front matter PII: S1359-6446(02)02218-3

tackle questions relating to DNA and protein sequences in bioinformatics, predicting 3D proteins structures in Folding@home (http://www.foldingathome.org), and seeking HIV-1 protease inhibitors in FightAID$@home (http://www.fightaidsathome.org), which uses AutoDock for VS and the Entropia software for networking (http://www.entropia.com).

Computational architecture

Although there are a vast number of networked PCs, there are some significant constraints surrounding memory usage, data transfer and algorithms that need to be considered. Within most companies, PCs have 24 hour access to the Internet, usually through a gateway, firewall and sometimes a proxy server. Compared with home and academic PCs, these are often faster and better configured. Consequently, from the researcher's perspective, these are the PCs of choice, but, unfortunately, most large corporations are unduly sensitive about use of these machines for the benefit of third parties. For home users, Internet access time is typically only minutes per day through a modem with transfer speeds up to 56 Kbaud (approximately 5.6 Kbytes s^{-1}). Software that can run productively for many hours without accessing the network and with minimal data transfer requirements is the best way to make use of the millions of PCs that are otherwise idle. Such parallel processing, when communication is occasional and there are constraints on the amount of data transferred, is sometimes termed 'coarse grained' parallel processing.

It is inevitable that application software needs to be rewritten or at least adapted to work in this fashion. Furthermore, the specialist programming tools adopted for parallel processing, with shared memory or high band-width connected processors, are obviously inappropriate when using very large numbers of occasionally networked PCs. The use of UNIX servers with typically 64 processors, such as those provided by Silicon Graphics (Mountain View, CA, USA; http://www.sgi.com), is already widespread within the pharmaceutical industry. Generally, it is easier to adapt software to use a small number of processors in parallel compared with the very different task of harnessing the potential of thousands of PCs on an intranet, or millions of PCs on the Internet. If the widespread use of multiprocessor servers is indicative of a contribution to the drug discovery process, then perhaps taking on orders-of-magnitude more processing power will revolutionize discovery timescales and costs.

Current implementations of Internet computing use a single server that dispatches queued tasks (or jobs) sequentially to clients and then receives the results. In an intranet environment, it is relatively simple for a programmer to use a shared network drive, and a file that serves as queue, with a locking mechanism to inhibit concurrent access to the queue to support up to 100 PCs. The THINK software can support this type of approach when the information read from the queue includes the input data file. However, on the Internet, remote access to information on a disk requires linking networking client software both to the application and to a file transfer protocol (FTP) server or similar. Although receiving 10–20 Kbytes of data for processing does not result in excessive PC-client connection time, supporting hundreds of thousands of PCs receiving data daily can exceed the capacity of a typical web server. This effectively limits the amount of data that can be processed per day or the numbers of PC clients that can participate in a single project.

For the application to be non-intrusive on the client's PC, it is important that it either stops completely when the user interacts with it (like a normal screen saver) or uses so little memory that it does not noticeably impair the performance of the PC. If this is not the case, users will quickly uninstall the software. In practice, it is necessary for the distributed software to be in at least two components: a permanently active agent that communicates with the server when necessary (the networking software), and the computational application. The permanently active agent is necessary to make or detect a network connection (when one is made on a home PC by the user dialing), send the results, and receive the data for the next job in a similar way to automatically using the FTP utility for transferring files. Such programs can be quite small (less than 32 Kbytes). Provided the computational application restarts like a screen saver when the PC is idle, but continues processing the data from where it stopped (or the last convenient checkpoint), limits on the memory usage are not significant. However, use of large amounts of disk space (that is, more than 10% of the disk capacity) is often problematic.

In addition, there are certain operational constraints that are inherent when using Internet computing. The best approach encrypts data and results files, in an attempt to prevent the clients from simulating results (perhaps motivated by improving their chances of winning prizes when these are offered), and compares results from two or more repetitions of the calculations. Such replication also addresses the inevitable failures of some clients to return results when, for instance, they discontinue participating in the project.

Applications

The choice of drug discovery project is never easy and for computational methods to be of most use it is necessary to model the interactions of small molecules and the proteins with which they interact. This implies that the 3D structure of the human protein target must be known. In practice, it is often desirable to use a similar mammalian protein for an *in vivo* assay, to understand as much as possible about the biochemical role of the protein and therefore enable estimation of selectivity and side-effects. Bioinformatics is still

immature with significant on-going development, and there is no consensus about what are the most appropriate questions to ask or which software to use. Nonetheless, many individuals wish to identify amino acid or DNA sequences similar to those of interest to gain an insight into other members of the target-protein family. The solution is intrinsically suited to parallel or distributed computing because it is appropriate to search for sequence similarity rather than sequence identity. Two well-known programs are BLAST (Basic Local Alignment Search Tools) and HMMER, which takes its name from Hidden Markov Modelling, an algorithm that originated in speech-pattern recognition. These programs can benefit from distributed computing simply by splitting the set of sequences into subsets for each of the processors.

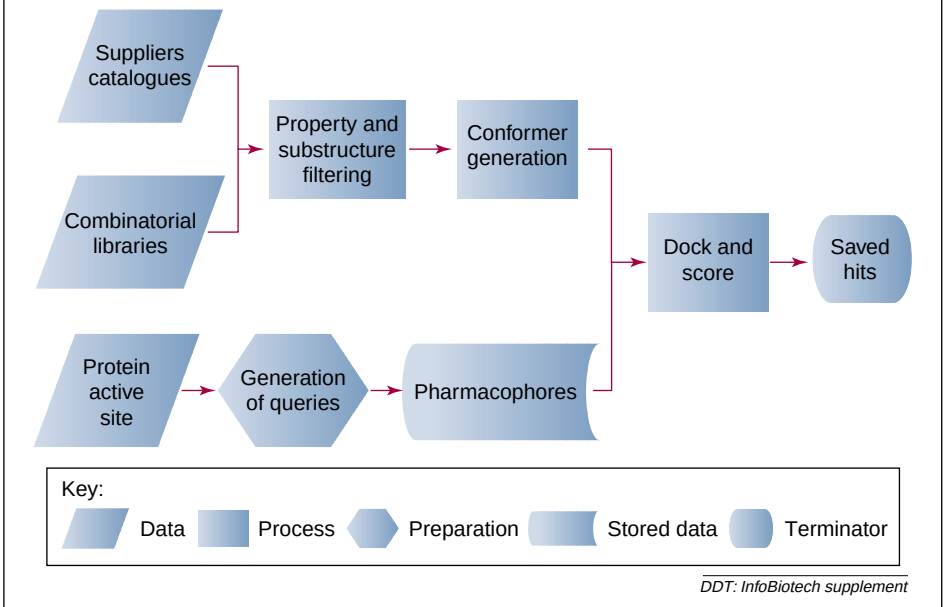

Figure 1. Flowchart of CAN-DDO virtual screening project

Probably one of the largest outstanding problems in computational molecular biophysics, is predicting the 3D structure of proteins – normally referred to as protein folding. If achieved, this would predict which of the many shapes or conformations (defined in terms of backbone and side-chain torsion angles) a protein would assume. Although it could be argued that the free-energy contributions arising from the enthalpy and entropy contributions are understood, it is computationally prohibitive to either generate all possible protein conformations or simulate protein motion for long enough to observe the whole of the folding process using molecular dynamics. As a consequence, the underlying science is technically unproven for the folding of all except small peptides. In the Folding@Home project, a compromise is reached whereby several possible trajectories are explored by simulation and the end-point of the most promising progress is used as the starting-point to generate the next set of trajectories. In practice, this approach can make only limited use of the parallel processing architecture inherent with Internet computing.

Dockcrunch [9] is one of the few published applications of parallel technology for drug discovery in the field of VS. In this project, some 1.1 million molecules were screened over six days for docking into the estrogen receptor on a 64 CPU Silicon Graphics server. Parts of the process, such as eliminating molecules that are non drug-like [10] and selecting which of the hits to make and test, are recognizable subtasks. Depending on the methods used, these can be quite time consuming. Of the 37 non-steroidal compounds finally selected for testing, 21 were found to have activity below 300 nM, and two of these were extremely active with activities below 10 nM. Today, there are two large-scale VS projects that use home computers connected to the Internet.

FightAids@Home is especially challenging because HIV mutates very rapidly (up to a billion mutants per day in a single individual) giving rise to mutant forms that are resistant to certain drugs. To prevent the emergence of drug resistance in a patient with HIV, it is necessary to inhibit a series of HIV mutants with many drugs. The FightAIDS@Home system is being used in several ways: (1) to screen against wild-type HIV protease using molecules that are available from the NCI (National Cancer Institute, Bethesda, MD, USA); (2) to test clinically approved inhibitors against virtual mutants of HIV-1 protease; and (3) to apply the principles of co-evolution to design novel 'master key' inhibitors capable of universally inhibiting a competing population of mutating HIV proteases.

The CAN-DDO project is much more ambitious, aiming to evaluate 3.5 billion molecules as inhibitors of 16 protein targets relevant to cancer therapy. Since the start of the project in April 2001, eight targets have been processed and the project is expected to complete in 2002. The THINK software uses a pharmacophore approach and full conformational searches, similar to that used in Chem-X [5] (Fig. 1). Like some other VS software, THINK ranks the hits using the ChemScore [11] function developed by Proteeus [Macclesfield, UK (now Tularik, San Francisco, CA, USA; http://www.tularik.com)], which provides an estimate of the free energy of binding for the conformation of each bound molecule. In VS, it is usual to assume that the

side-chains in the receptor site are stationary. In the pharmacophore approach, some allowance for flexibility can be accommodated in tolerances in the positions of the interaction centres that are matched to those in the small-molecule ligands.

Like the FightAids@home project, THINK uses a genetic-like algorithm to mutate a starting molecule into derivatives. Thirty-five million molecules were collected from small molecules that are commercially available for screening and from 23 combinatorial chemistry libraries. These starting molecules were the results of pre-filtering to eliminate those that fail to exhibit drug-like properties or contain undesirable functional groups (e.g. those that are toxic, reactive or easily metabolized). Such drug-like filters are also applied to the derivatives that are generated. Thus, by generating 100 drug-like derivatives from 35 million molecules, approximately 3.5 billion molecules are constructed for VS. THINK integrates within a single application filtering, the creation of 3D coordinates from connection tables, the generation of conformers, docking, and scoring. This is convenient for VS and perhaps essential for effective distributed processing as well as for avoiding the integration issues encountered by other applications [12]. However, researchers who prefer to pick 'best-of-breed' algorithms, or programs that use different approximations, must also face the significant development costs of associated architecture constraints and post-processing of the results to eliminate undesirable hits.

Although it is outside the scope of this review to discuss in detail the interim results of any of these on-going projects, the corresponding websites give further information. However, it is relevant to appreciate the numbers of hits generated and the scale of post-processing that might be required. The numbers of crude hits can be very large, and for CAN-DDO presently ranges from 7500 to 2.4 million (a hit rate of less than 0.001%). This is usually significantly reduced by refining the torsion angles and the x,y,z orientation and position of the ligand in the receptor site, and only accepting hits that achieve a lower threshold score. In addition, in this project, the generation of derivatives enables families of hits to be readily identified. It should also be appreciated that the time for post-processing can easily exceed 100 days on 2 GHz-PCs per protein target, and it might be of little value to increase the numbers of molecules screened without developments in post-processing techniques for the hits. Work is underway to validate experimentally the binding predictions and identify the percentage of false-positive hits.

Achieving its potential?

Currently, DNA and protein-sequence searching can be performed using supercomputers, but ultra-high-throughput VS requires much more computing power such as that available via the Internet. The limited network bandwidth for home PCs means that predictions of protein folding using molecular dynamics are unable to use the potential of these PCs. Perhaps this will remain unaddressed until better algorithms or hardware become available.

VS applications approach the limits of the resources available for Internet computing: AutoDock requires at least 96 Mb of RAM to run on the client PC and the downloads approach 200 Kbytes of data. In the CAN-DDO project, a cut-down version of THINK is used that uses much less memory than AutoDock, and the amount of data transferred per job is only 20 Kbytes because no coordinates of the small molecules are transferred. As a consequence, CAN-DDO is able to use many more client PCs and evaluate the potential of many more molecules.

Before starting a project, a sponsor might want to consider patent and security issues. Can the hits be identified on client PCs and would clients have rights to patent hits or claim payment for discoveries? Are the sponsor's rights in a discovery patent (usually filed when biological activity is confirmed experimentally) weakened when the drug was found by a third party? What level of security is appropriate to prevent malicious clients returning incorrect results to the server and extracting hits or virtual collections of molecules? Although the client agrees to certain terms and conditions before participation, the legal issues in a commercial setting of a pharmaceutical company might cause some to hesitate or limit distributed-computing in drug discovery to the intranet rather than the Internet.

It has been previously observed that the pharmaceutical industry has 'a history of entrenched practices and resistance to innovation in drug development' and that improving productivity by adopting new technology 'will require fundamental change in the way that pharmaceutical drug discovery is conducted' [13]. These comments are probably unfair to small biopharmaceutical companies, but many chemists are perhaps unduly concerned about missing hits even though the numbers of hits from VS is often so large that it is not obvious how to review and select a subset for further testing. The evaluation of many derivatives can distinguish families of hits from dead-end leads, which is one of the advantages of considering the derivative-series before deciding which hits are valuable leads. Unfortunately, current technology gives a significant percentage of false positives and more research is required to establish whether these arise because of thermodynamic or kinetic issues. This means that there will continue to be a need for experiment and serendipity – but the odds have certainly been greatly improved.

References

1 Kuntz, I.D. *et al.* (1982) A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* 161, 269–288

2 Morris, G.M. *et al.* (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comp.-Aided. Mol. Des.* 10, 293–304

3 Li, J. *et al.* (1998) Targeted molecular diversity in drug discovery: integration of structure-based design and combinatorial chemistry. *Drug Discov.Today* 3, 105–112

4 Rarey, M. *et al*. (1997) Multiple automatic base selection: protein–ligand docking based on incremental construction without manual intervention. *J. Comp.-Aided Mol. Des.*11, 369–384

5 Murray, C.M. and Cato, S.J. (1999) Design of Libraries to Explore Receptor Sites. *J. Chem. Inf. Comput. Sci.* 39, 46–50

6 Miller, M.D. *et al*. (1994) FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comp.-Aided Mol. Des.* 8, 153–174

7 Warr, W. (2001) *Virtual High-Throughput Screening: Computational Tools for Drug Discovery and Design in Spectrum Life Sciences*, Decision Resources, MA, USA (http://www.decisionresourses.com)

8 Hand, L. (2001) Computing for cancer research. *The Scientist* 15, 1–5

9 Waskowycz, B. *et al.* (2001) Receptor-based virtual screening of very large chemical datasets. In *Rational Approaches to Drug Design* (Holtje, H-D. and Sipple, W., eds), pp. 372–381, Prous Science, Barcelona

10 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25

11 Murray, C.W. *et al.* (1998) Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand-receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comp.-Aided Mol. Des.* 12, 503–519

12 Walters, W.P. *et al.* (1998) Virtual screening – an overview. *Drug Discov.Today* 3, 160–178

13 Beresford, A.P. *et al*. (2002) The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov.Today* 7, 108–116