

TOWARDS NEW DRUG THERAPIES: Computer assisted design of peptide- based protease inhibitors. Elastase and emphysema.

J. Crisp, S.R. Smith, E.L.R. Compton, D. Kidger*,
R. McConnell*, A.R. Clarke & R.B. Sessions

Department of Biochemistry, School of Medical Sciences,
University Walk, Bristol BS8 1TD, UK

*ClearSpeed Technology, 3110 Great Western Court,
Hunts Ground Road, Bristol, BS34 8HP, UK
www.clearspeed.com

Abstract

Molecular docking programs are widely used in drug discovery and there is intense research into methods of rapidly predicting binding energies. We present here an approach which uses an empirical free energy forcefield based on pairwise atomic interactions between ligand and receptor. This method is intrinsically more computationally demanding than more traditional scoring functions. The method is implemented in the program BUDE and

used to design peptide libraries in the search for inhibitors to the protease, human elastase as therapeutics or lead compounds for controlling the effects of lung tissue degeneration in emphysema. By using the extraordinary performance/power characteristics of the ClearSpeed CSX600 we show how this process can be massively accelerated.

Proteases, a brief background to structure and function

Proteases are enzymes (catalytic proteins) that catalyse the cleavage of peptide bonds. Since all proteins are composed of linear chains of α -aminoacids, joined by such peptide bonds, the proteases have the potential to self-destruct via a process known as autolysis (self-cleavage). There are two factors that reduce the rate of autolysis in native proteases. Firstly, as in all natural proteins the polypeptide chain is built up from a precise sequence of the 20 different naturally occurring aminoacids which folds to a unique three-dimensional structure. This structure provides both the catalytic machinery (a polypeptide binding site and precisely juxtaposed chemical groups for catalysis) and a degree of protection to its own polypeptide chain by burial in the globular structure and some conformational rigidity to its surface exposed loops. Secondly, nature designs proteases to cut polypeptide backbones in precise positions by recognising and binding particular sequences of amino acids. The two properties, molecular recognition according to sequence and stability to proteolysis (cleavage) conferred by conformational rigidity, will be exploited in the design of small peptide-based protease inhibitors as described below.

Proteases, a brief background to their biological importance

Recent studies indicate that 1-5% of genomes code for proteases. This observation reinforces the key role that proteases play in many biological processes and diseases including cell signalling, pro-enzyme maturation, viral infection, blood clotting, hypertension and Alzheimer's disease, to name a few. Many protease inhibitors are currently marketed as drugs. These may target pathogens, for example protease inhibitors that block viral replication are a key component of the anti-HIV triple therapy, or human proteases, for example ACE inhibitors block the action of the protease angiotensin-converting enzyme and are used for the control of blood pressure.

The target of this work

Emphysema is a chronic obstructive pulmonary disease (COPD). The symptoms are breathlessness and shortness of breath caused by a reduction in the elasticity of lung tissue, compromising the ability of the lungs to deflate appropriately during breathing. Environmental factors such as smoke inhalation causes the secretion of the protease elastase from lung cells. This elastase breaks down elastin in lung tissue which is itself a protein component responsible for lung tissue elasticity¹. Hence elastase inhibitors are expected to prevent further elastin degradation and halt the progression of the disease^{2,3}.

Small peptide-based inhibitors, a brief background

Peptides are simply short lengths of natural polypeptide and are typical substrates and products of proteases. One small peptide (14 amino acid residue, SFTI-1) was isolated from sunflower seeds and found to be a potent inhibitor of the protease trypsin⁴ with a K_d around 1×10^{-10} M. Solving the three dimensional structure of the complex between trypsin and SFTI-1 showed that the inhibitor was bound into the active site of the enzyme, blocking its ability to cleave normal substrate peptides and proteins. The reason the inhibitor itself was not cleaved was due to the rigid and constrained structure of the peptide⁴. This rigidity and constraint is conferred by a cyclic backbone structure, further braced by a disulphide bridge between two cysteine amino acid residues and cross-ring hydrogen-bonding (figure 1a).

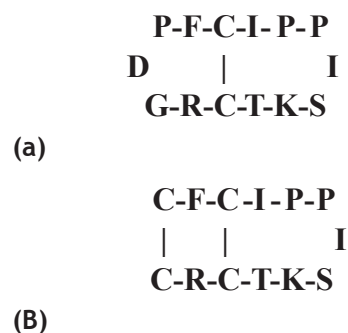


FIGURE 1. (A) THE AMINO ACID SEQUENCE OF SFTI-1 SHOWING THE CYCLIC BACKBONE AND SINGLE DISULPHIDE BRIDGE (B) THE AMINO ACID SEQUENCE OF THE READILY SYNTHESISED ANALOGUE.

Redesign of the inhibitor for rapid library synthesis

Molecular modelling indicated that backbone cyclisation of SFTI-1 could be replaced by a second disulphide bridge without seriously interfering with binding to the trypsin active site (figure 1b). Hence the corresponding linear peptide was made by standard peptide synthesis, and cyclised via disulphide bond formation by exposure to atmospheric oxygen. This compound had a K_d of 5×10^{-10} M, hence binding only slightly less tightly than the native inhibitor. The binding mode to trypsin was shown by crystallography to be the same as that of SFTI-1. Inspection of the crystal structures identified five amino acid residue positions on the peptide that could be altered to affect the interaction of the peptide with proteins while retaining those residues responsible for resistance to degradation by proteases (figure 2). Since each of the five positions could be occupied by one of 20 amino acids the total number of possible compounds that could be synthesised is $20^5 = 3.2 \times 10^6$.

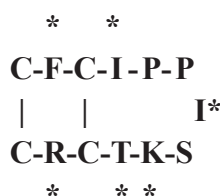


FIGURE 2. ASTERISKS SHOW THE FIVE RESIDUE POSITIONS AVAILABLE FOR ALTERATION.

Computer assisted design of focussed peptide libraries

It is clearly impractical to individually synthesise and test the inhibitory properties of each possible peptide sequence (similar to the total number of compounds available to the worlds pharmaceutical companies for testing against various targets via robotic high-throughput screening). Likewise, mixed synthesis methods could easily be used to make all 3.2×10^6 compounds in one mixture but the amount of each compound would be vanishingly small, impossible to test or isolate. What is required is a half-way house. For technical reasons related to solubility and

concentration, it transpires that a library containing about 100 different peptide sequences in one pot is the maximum convenient size for testing and identification of a single (or series) of inhibitors. Such a library is easily prepared by mixed synthesis.

An experienced molecular modeller can use molecular graphics methods to generate an initial docking position (pose) of a generic cyclic peptide (eg containing alanine at each of the 5 variable positions). Since evolution has selected the 20 natural amino acids to cover a wide range of chemical diversity, the individual amino acids can be grouped into a variety of types that include large, small, hydrophilic, hydrophobic, positively charged, negatively charged. Hence the molecular modeller can also make predictions of what type of amino acid would be best matched to the particular environment surrounding the five variable positions in the initial pose. The modeller strives to choose an average of 4 or fewer candidate amino acids for each position, yielding a virtual library of peptide sequences of $4^5 = 1024$ members or less. It is the purpose of the computer algorithm described here to bridge that gap and determine which 10% of the virtual library is the best choice for actual synthesis and testing.

BUDE (Bristol University Docking Engine)

The overall algorithm is composed of the following elements:

1. The user defines a discrete search space (a 6-D grid) around the initial ligand (peptide) pose
2. The fitness of a pose is evaluated by a novel atom-atom based empirical free energy forcefield.
3. The grid positions may be evaluated exhaustively or by using a genetic-algorithm-like Monte Carlo search method (EMC)⁵.
4. Ligand flexibility is treated by docking different conformations of the peptide.
5. Docking many ligands is a trivially parallel problem.

1. - The search space: This is simply a rectilinear grid defined by user-supplied increments in Angstroms in the x y and z Cartesian axes and user-supplied increments of rotation in each of these axes. A typical search space (used in this case study) is comprised of translations -3, -2, -1, 0, 1, 2, 3 Angstrom in each axis and rotations of -30, -20, -10, 0, 10, 20, 30 degrees in each axis. This yields $76 = 117649$ poses.

2. - The forcefield: Docking programs typically use simple heuristic scoring functions that are extremely fast to calculate but do not necessarily provide sufficient accuracy to locate correct poses and are even less likely to reproduce real binding affinities. At the other end of the scale, extremely computationally intensive methods for successfully determining relative free energies of binding of similar ligands to a protein may be used, such as free energy perturbation or thermodynamic integration (ref jorgenson), and cannot currently be used for screening many compounds. At first sight, the middle ground would appear to belong to methods like molecular mechanics minimisation and continuum electrostatics. However, such methods do not account for entropy and desolvation effects and are poor at reproducing binding energies. The major focus of methods development in this research is to design and refine an empirical free energy forcefield based on pairwise atomic interaction functions (by analogy with the development of molecular mechanics pairwise potential energy functions, a process that has been ongoing since the 1960's).

The starting point for the Bude forcefield (BuFF) is the Raft forcefield developed by us for fold prediction of peptides and small proteins⁵. The solvation-based soft potentials of Raft are projected onto the all-atom (heavy atoms only) description of structure used in Bude and hydrogen bonding and charge-charge interactions are modelled by modified Coulombic interactions.

3. - The search method: Within the above formalism, any pose can be described by a six integer grid address or pose descriptor. Mutation of this pose descriptor is a simple way to implement a GA-like EMC minimisation protocol⁵. We find that searching about 5% of the grid space this way is sufficient to find either the global energy minimum or a pose and energy adequately close for our purposes.

4. - Ligand flexibility: Most of the 20 natural amino acids have flexibility in their sidechains due to rotation about their single bonds. However, it is well known that such conformational variability can be approximated by using a library of different conformers for each sidechain (rotamers). Hence each aminoacid has a set of conformations available, numbering between 1 and 13 depending on its type. The Bude user can configure the program to use these rotamers and dock each ligand conformation separately, although a little care has to be taken to avoid a combinatorial explosion.

5. - Many ligands: In this case study a library of 576 different peptide sequences was generated and each docked as a separate Bude job. Shell scripting is used to address this problem of trivial parallelisation. Each sequence generates between 80 and 30,420 conformations to be docked, depending on the number of rotamers associated with each aminoacid in the peptide. In total, 1,966,272 docking operations are performed. Since each docking operation searches some 5% of the grid (4225 poses), the energy of 8,307,499,200 (i.e. over 8 billion) poses must be calculated to evaluate the whole library.

ClearSpeed acceleration

The “currency” of the EMC minimisation is the pose descriptor, it is the task of the docking engine to translate that pose descriptor into a pose energy. The Bude source code is about 5,000 lines of FORTRAN. Profiling the code shows that as expected > 99% of the execution time is spent in the energy calculation routine which is about 500 lines of code. Only this routine needs to be accelerated on the ClearSpeed Advance accelerator but for convenience the geometry transformations are also performed on the card. Before the EMC begins, the initial coordinates of the protein (elastase) and the ligand (cyclic peptide) are copied to the Advance accelerator's on board DRAM. When the EMC requires the energies of a set of pose descriptors the host program translates these into a set of transformation matrices. This set is copied to the Advance accelerator, the transformations applied and the energy calculated, and the results copied back to the host process. The time to calculate the geometry and energy of a single pose on an Advance accelerator board is about 1 ms. Consequently, the 8 billion poses for the elastase peptide library would take about 90 days. Due to the trivial parallelism afforded by the 576 independent amino acid sequences this problem essentially scales perfectly such that a whole peptide library calculation on the 144 boards in the configuration of twelve node ClearSpeed Accelerated TeraScale System (CATS), will take about 15 hrs.

Future developments

1. Replace the exhaustive evaluation of every peptide (ligand) conformation with an EMC minimisation of the rotamers during a single docking of a given cyclic peptide sequence.
2. Include flexibility of protein (receptor) sidechains in the binding site and treat these as in point 1.
3. Further tune the ClearSpeed Advance accelerator code
4. Add dynamic load balancing
5. Implement other mutation methods
6. Replace the shell scripting harness by an MPI harness.

CONCLUSION

We have shown how a substantial acceleration of a molecular docking program using a “next generation” empirical free-energy forcefield can be readily achieved using ClearSpeed Advance acceleration. Using 12 boards in a CAT system we obtain an acceleration compared with the host quad-core 3 GHz Xeon host of 10.2, allowing studies such as that described here to be performed in days rather than weeks. Of even greater significance is the performance per watt, considering that the cost of large computer clusters is increasingly shifting from the capital expense of hardware purchase to the provision of cooling and power via dedicated machine rooms. In terms of performance per watt, we see at least a 5 fold increase over state-of-the art quad-core conventional processors. In the context of molecular docking this may be exploited either by performing the same amount of searching in a smaller time or searching a larger area of chemical space in the same overall time, entirely as the user wishes.

References

1. R.D. Hautamaki, D.K. Kobayashi, R.M. Senior, S.D. Shapiro, *Science* **277** 2002-2004 (1997)
2. S.D. Shapiro, N.M. Goldstein, A. McGarry Haughton, D.K. Kobayashi, D. Kelly, A. Belaouaj, *Am. J. Path.* **163** 2329-2335 (2003)
3. P.J. Barnes, *Curr. Drug Targets Inflamm. Allerg.* **4** 675-683 (2005)
4. S. Lockett, R.S. Garcia, J.J. Barker, A.V. Konarev, P.R. Shewry, A.R. Clarke, R.L. Brady, *J. Mol. Biol.* **290** 525-533 (1999)
5. N. Gibbs, A.R. Clarke, R.B. Sessions, *Proteins* **43** 186-202 (2001)

Copyright 2007 ClearSpeed Technology plc (“ClearSpeed”).

All rights reserved.

All information in this document is provided only as general information in connection with ClearSpeed products. Except as provided in ClearSpeed’s terms and conditions of sale for such products, ClearSpeed assumes no liability whatsoever, and ClearSpeed disclaims any express or implied warranty relating to sale and/or use of ClearSpeed products, including liability or warranties relating to fitness for a particular purpose, merchantability, or infringement of any patent, copyright, or other intellectual property right. ClearSpeed may make changes to specifications, product descriptions, and plans at any time, without notice.

ClearSpeed, ClearConnect and Advance are trademarks or registered trademarks of ClearSpeed Technology plc or its group companies. All other marks are the property of their respective owners.

V1 0711