UPTEC X 01 013
FEB 2001

ISSN 1401-2138

KATARINA JANSSON

A pilot study for virtual screening: finding inhibitors of hen egg-white lysozyme

Master’s degree project

Molecular Biotechnology Programme
Uppsala University School of Engineering

UPTEC X 01 013 | **Date of issue 2001-02**

Author

Katarina Jansson

Title (English)

A pilot study for virtual screening: finding inhibitors of hen egg-white lysozyme

Title (Swedish)

Abstract

Virtual screening (VS) is a computer-based method to search for new drugs or protein ligands. In this work some of the VS methods available and computer programs for this purpose have been tested. Hen egg-white lysozyme (HEWL) was chosen as the model receptor system and the molecule database chosen for screening was the Available Chemical Directory (ACD) database. An iterative protocol using docking simulations combined with diversity and similarity selections, and a flexible search where the database was explored using a query defined by interaction properties gave promising results. The hits obtained by both methods cluster in a small area of the descriptor-space defined by the molecular distribution of hydrophobicity and partial charges. This could be interpreted as the discovery of a new "activity island" for inhibitors of lysozyme and probably even other glycosyl hydrolases.

Keywords

Virtual screening, molecular modelling, docking, diversity selection, similarity selection, flexible query search, lysozyme, inhibitor

Supervisors

Dr. Enrique Carredano
Polymer and Surface Chemistry, Amersham Pharmacia Biotech

Examiner

Dr. Gerard Kleywegt
Dept. of Cell and Molecular Biology, Uppsala University

Project name	Sponsors
Language **English**	Security
ISSN 1401-2138	Classification
Supplementary bibliographical information	Pages **46**

Biology Education Centre Biomedical Center Husargatan 3 Uppsala
Box 592 S-75124 Uppsala Tel +46 (0)18 4710000 Fax +46 (0)18 555217

A pilot study for virtual screening: finding inhibitors of hen egg-white lysozyme

Katarina Jansson

Sammanfattning

Virtuell screening kan förklaras som ett sätt att hitta nya proteinbindare m.h.a. olika datorprogram och proteinstrukturer. I denna studie användes lysosym från hönsäggvita som modellprotein och en stor molekyldatabas, Available Chemical Directory (ACD), för att testa virtuell screening. Flera olika metoder prövades och två av dessa gav lovande resultat. Dels en iterativ metod där man utgick från en utspridd delmängd av den stora ACD databasen och dockade dessa molekyler i den aktiva ytan på lysosym. De molekyler som rankades bäst i dockningen användes sedan för att söka efter liknande molekyler i databasen varefter även dessa molekyler dockades, denna procedur upprepades fyra gånger. Den andra metoden utgick från en definition av några viktiga interaktioner mellan lysosym och dess ligand, denna definition användes för att söka i ACD databasen efter nya molekyler som hade motsvarande bindningsmöjligheter. Även dessa molekyler dockades sedan i den aktiva ytan. Två ”aktivitetsöar” för lysosym, definierade av distributionen av partiella laddningar och hydrofobicitet på molekylerna, hittades varav den ena var helt ny och innehöll huvudsakligen molekyler med aromatiska ringar medan den andra innehöll sockerliknande molekyler och även de naturliga bindarna till lysosym.

Examensarbete 20 p i Molekylär bioteknikprogrammet

Uppsala universitet februari 2001-02-12

1. Introduction

1.1 Virtual Screening

Virtual screening (VS) may be defined as the automatic evaluation of a large database of compounds using a computer. The goal of VS should be that the selected compounds meet certain requirements which make them good candidates for binding a protein receptor. VS works best in an information-rich environment. This information may include knowledge about compounds that interact with the receptor, knowledge about the receptor structure and receptor-interactions and even knowledge about drugs in general[1] (Walters *et al*. 1998).

Since the 1980's structure-based methods have been used in the development of several drugs on the market today. Some examples are listed in Table 1.

Drug	**Target**	**Reference**
Cozaar	Angiotensin II(AII) receptor	Dunica *et al*., 1990; Dunica *et al*., 1992
Aricept	Acetylcholinesterase	Kawakami *et al*., 1996
Teveten	Angiotensin II converting enzyme	Weinstock *et al*., 1991; Keenan *et al*., 1993
Zomig	Serotonin receptor	Glen *et al*., 1995
Viagra	CGMP phosphodiesterase	Terrett *et al*., 1996
Crixivan	HIV-1 and HIV-2 protease	Dorsey *et al*., 1994; Holloway *et al*., 1995
Viracept	HIV-1 protease	Kaldor *et al*., 1997

Table 1 Examples of drugs on the market today developed using structure-based methods.

When the goal is to search for new ligand molecules in virtual libraries, VS may be divided into five different types[2] listed below in order of increasing complexity and knowledge about the target system as well as increasing computation-time requirements.

[1] One form of VS is compound filtering. By discriminating against compounds violating the so-called Lipinski rules limiting LogP, molecular weight and the number of H-bond donors and acceptors drug-design modellers ensure the drug-like character of the molecules. This aspect will not be taken into account in this work.

[2] It is also possible to refine and optimise structures by building small, directed virtual libraries constructed for a specific target. Experimentally obtained information may be used in VS in the form of quantitative structure activity relationships (QSAR's). However this procedure of 'lead optimisation' is outside the scope of this work.

i) Even without using any information about the target or known substrates or inhibitors, the odds of finding a ligand by chance can be improved in computer-based ligand design. A representative subset of compounds may be selected from a large database without losing the diversity of the database (Pötter and Matter, 1998). The *dissimilarity* selection is dependent on the particular algorithm and the structural descriptors used.

ii) There are efficient *similarity* screening methods using molecular descriptors. Similarity searches are usually based on already known active compounds, *e.g*. natural substrates or inhibitors. Depending on the descriptors used, the similarity search results in a list of similar compounds (neighbours) in terms of physico-chemical properties or two-dimensional structure.

iii) Molecules may be *superimposed* onto the active compounds to find new ligands with similar molecular properties. This is done using efficient algorithms, which allow variation of the rotational angles and explore the binding properties of the molecules in the database in order to find the best match.

iv) In *three-dimensional searches* the starting point is not a known ligand molecule by itself, but the crystal structure of a complex of a ligand and the receptor or the receptor alone. A pharmacophore description of the target can be defined by studying the shape and the arrangement of possible interaction spots within the active site. The pharmacophore may then be used as a query to search for leads in a three-dimensional database of compounds. Some programs like UNITY (Tripos Inc.) include the option of *flexible searches* allowing various values for the torsion angles of the small molecules in the database.

v) The recent development of faster and more reliable *virtual docking* algorithms have made it possible to dock not only the hits from the 3D-search into the active site of the target, in order to evaluate those hits further, but even in some cases the entire database or a representative subset. The docked compounds are ranked according to, for example, docking energies. Docking simulations generally neglect entropic and solvent effects. These and other approximations result in severe limitations of the reliability of docking energies (see for instance Bissantz *et al.* 2000 for a review about docking scores and their limitations). In a study using the fast docking program FlexX (Rarey *et al*., 1996), 10 out of 19 complex crystal structures were reproduced accurately (rmsd < 1.5 Å) by the best ranked docking conformation.

However, for only 5 of the 10 was the predicted binding energy comparable (within 2 kJ/mol) to the experimentally obtained energy of binding[3].

Finally, the different methods *i*-*v* may be combined to improve their efficiency. Method *iii* may be thought of as a more sophisticated variant of *ii*, since both methods are purely ligand-based. Similarly, method *v* is more sophisticated than method *iv*, but both methods are mainly receptor-based. The simpler methods have the advantage of being faster, enabling the use of a larger database. They could be used in an early phase whereas the more complicated methods could be used in a later phase of VS or for exploring a representative subset of the database as obtained from *i*. A disadvantage with receptor-based methods (*iv* and *v* above) is the necessity of having high-resolution crystal structures.

One recent example of a success story of VS has been presented (Gruneberg *et al*., 2001). Programs from the TRIPOS (Tripos Inc.) software package were used to obtain novel human carbonic anhydrase II inhibitors. The Available Chemical Directory (ACD) database (MDL Information Software Systems) was screened with the programs UNITY, FlexS (Rarey et al., 1999) and FlexX. 180000 compounds were screened with a 2D Query (minutes). This reduced the number to 57000 compounds and subsequently a 3D Query (4 days) gave 44000 compounds which were submitted to a superposition-screen (method *iii* above) using the program FlexS (12 days). Finally the 100 highest-scoring compounds were docked with FlexX (hours). After visual inspection a test set of 38 compounds was obtained. An IC50 assay later confirmed that 50% of these compounds had affinities in the μM range or higher, a few had sub-μM and three had sub-nM affinities.

The major advantages of VS are the possibility of greatly reducing the number of new compounds that must be synthesised and the possibility to cover a much larger structural property space with a virtual combinatorial database than with a combinatorial library created in the laboratory (Walters *et al.* 1998).

1.2 Goal and biotechnological importance of this work

Advances in computing technology, molecular modelling software, molecular biology techniques and an increasing availability of high-resolution X-ray structures of protein

[3] Efforts to calculate more accurate free energies of binding have been made (see for instance Hansson *et al.*, 1998), however these methods are not high throughput and were therefore not considered in this study.

targets have made VS feasible. VS technology has the potential to substantially increase the efficiency of the drug discovery process and structure-based ligand design in general. One of the challenging tasks which has arisen is the validation of these procedures by determining how close the findings obtained by VS protocols are to experimental results. When doing this, it is important to start with a generous scenario. The ideal receptor system should be rich in information and the compounds of the database should be readily available. For these reasons hen egg-white lysozyme (HEWL) was chosen as the model receptor for this study. The database chosen was the ACD which is essentially a catalogue over most of the compounds that can be purchased off the shelf from commercial vendors. The main purpose of this work was not to find inhibitors to HEWL, which may lack commercial value, but rather to test the feasibility of virtual screening (in particular methods *i*,*ii*,*iv* and *v* outlined above) and the computer programs used. The actual candidate-inhibitors obtained could then be purchased and experimentally tested by any suitable method. The experimental part of this validation study is beyond the scope of the present work. There may be another important spin-off of this study. Lysozyme is a very often-used model enzyme[4] and its role as a model enzyme in protein crystallography has been reviewed (Strynadka and James, 1996). To the author's knowledge no non-carbohydrate ligand to the active site of hen egg-white lysozyme has been found before. Therefore the results of an experimental screening of the candidate list obtained through this work may contribute to the knowledge base regarding molecular recognition not only for lysozyme but also for other glycosidases and enzymes in general.

1.3 Hen egg-white lysozyme

Hen egg-white lysozyme was the first enzyme and the third protein to have its structure determined crystallographically (Blake *et al.*, 1965, 1967a,b). Lysozyme catalyses the hydrolysis of β-(1-4) glycosidic linkages of various oligosaccharides, especially those of the bacterial cell wall. The active site of HEWL has the form of a long cleft and it is capable of binding six sugar residues simultaneously. Early modelling of polysaccharide substrates into the active site of HEWL showed six

subsites designated A to F (Blake *et al*., 1965; Johnson & Phillips, 1965). Substrate analogues, consisting of *N*-acetyl glucosamine (abbreviated GlcNAc or NAG) or alternating NAG and *N*-acetyl muramic acid (abbreviated MurNac or NAM) residues can be accommodated at all six subsites (Strynadka and James, 1996; Creighton, 1997).

1.3.1 Overall structure

Hen egg-white lysozyme contains 129 amino acid residues. The deep active site cleft divides the enzyme into two domains; one of them is almost entirely β-sheet structure, whereas the other contains both the N- and C-terminal segments and is mostly α-helical (Figure 1).

1.3.2 Active site

The catalytic residues Glu35 and Asp52 protrude into the active site from opposite sides of the active-site cleft (Strynadka and James, 1996). Even though lysozyme is a relatively rigid molecule the upper walls of the active site cleft (residues 45-50, 62, 73-75, 99-104) constitute a mobile portion of the protein (Strynadka and James, 1996; Steinrauf, 1998). This conformational flexibility allows the binding site to readily accommodate an incoming substrate (Figure 2). Especially the indole ring of Trp62 has been found to rotate towards the bound ligand, with a mean shift of the ring atoms of 1.0 Å (Cheetham *et al.* 1992). A more recent high-resolution unliganded structure (PDB code 1lks; Steinrauf, 1998) shows Trp62 with a completely different side-chain conformation (Figure 3).

[4] There are more than 600 lysozyme structures available in the Protein Data Bank (PDB), and almost 200 structures of hen egg-white lysozyme (October 2000), 66 of these with a resolution better than 2.0Å.

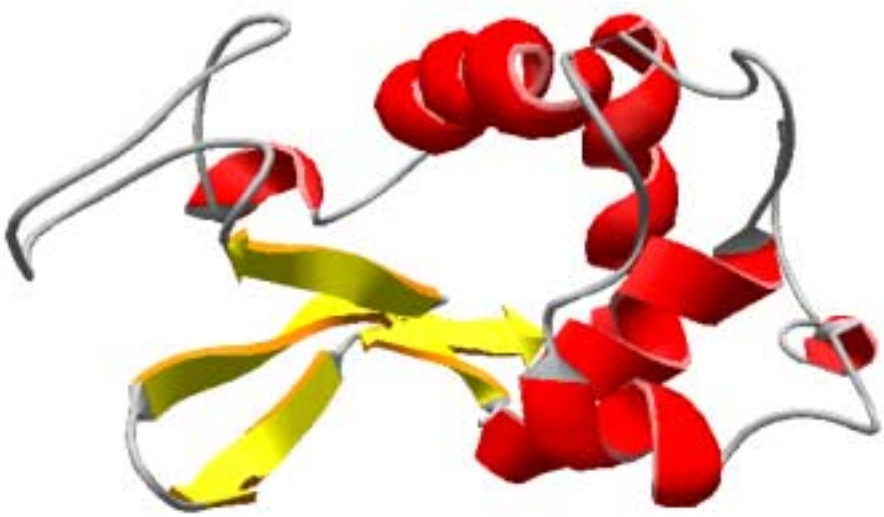


Figure 1 Overall structure of hen egg-white lysozyme with alpha helices shown in red and beta sheets in yellow (PDB structure 1lzb).

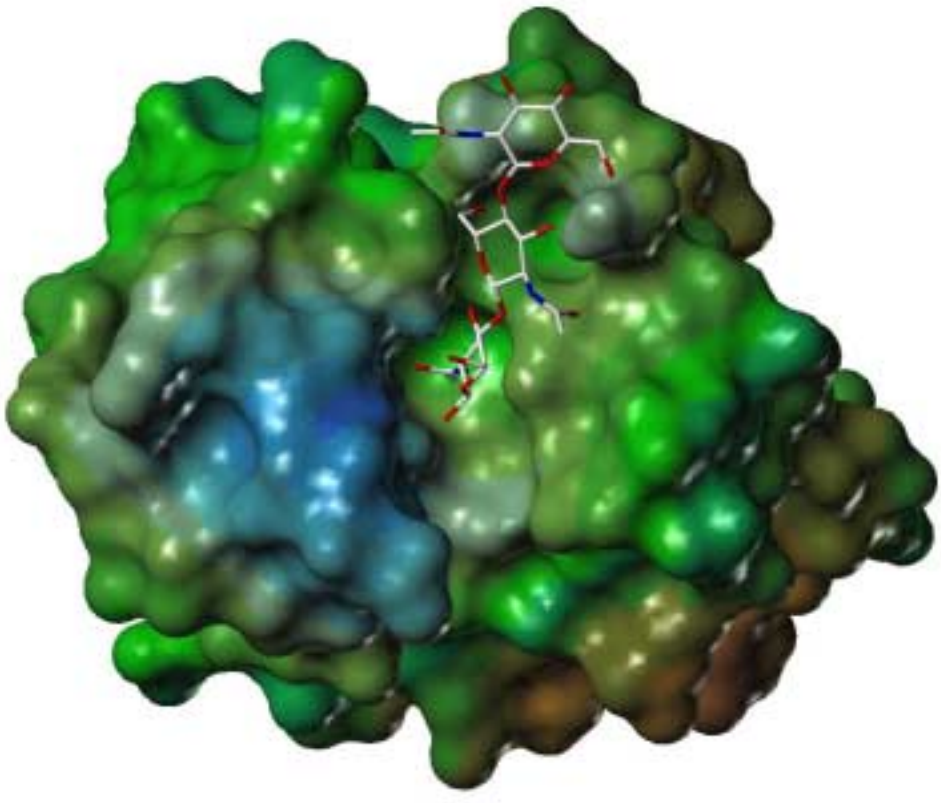


Figure 2 Hen egg-white lysozyme co-crystallised with the NAG_3 ligand (PDB code 1lzb). The ligand is positioned in the subsites A (the sugar ring at the top) to C (the sugar ring at the bottom). The surface is a fast Conolly surface colour-ramped according to lipophilicity, brown is most lipophilic and blue is least lipophilic.

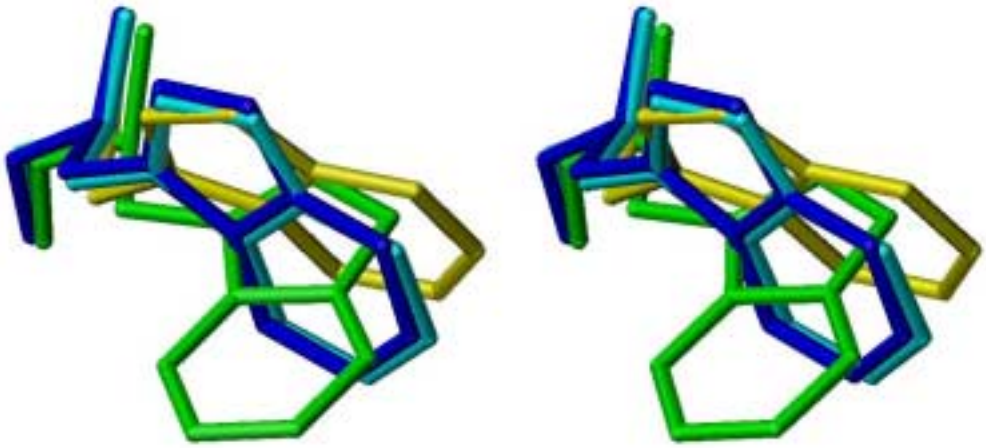


Figure 3 Stereo view. Different conforamtions of Trp62 in four high-resolution crystal structures both with ligand (PDB code 1lzb in cyan and PDB code 1hew in blue) and without ligand (PDB code 193l in yellow and PDB code 1lks in green)**.**

The active site, as mentioned above, can be divided into six subsites. The NAG_3 molecule used in many structure determinations, is accommodated in sites A to C. The sugar residue in site A is the least tightly bound of the three saccharide rings. There are two hydrogen-bonding possibilities in subsite A. One H-bond involves the hydroxyl oxygen (O6) in the sugar ring and either Asn103OD1 (PDB-structure 1hew) or Asn103ND2 (PDB-structure 1lzb). A second H-bond involves the nitrogen (N2) of the ligand *N*-acetyl group and one of the carboxyl oxygens (OD2) of Asp101 (PDB-structure 1lzb). This carboxyl oxygen may also hydrogen bond to the hydroxyl oxygen (O6) of the saccharide ring in subsite B. In subsite B, the NAG molecule is "stacked" with its apolar face towards the plane of the indole ring of Trp62. This stacking interaction is typical of the general features of protein-carbohydrate interactions and indicates a strongly hydrophobic interaction between the sugar and the protein (Cheetham *et al*., 1992). Subsite C has a total of six hydrogen bonds, including waters in a hydrogen-bonding network. There are four hydrogen bonds directly between the

enzyme and the sugar ring. The hydrophobic plane of Trp63 approaches the polar face of the saccharide ring at right angles allowing a hydrogen bond between the indole nitrogen and the carbonyl oxygen (O3) of the sugar. The indole nitrogen of Trp62 is also directed towards the saccharide ring making a hydrogen bond with the hydroxyl oxygen (O6). There are two hydrogen-bonding main chain atoms: Asn59N and Ala107O making H-bonds to the carbonyl oxygen (O7) and the nitrogen (N2) of the ligand *N*-acetyl group, respectively. The apolar face of the sugar ring in site C lies in a relatively hydrophobic environment with residues Ile98, Ala107, Trp108 and Val109 (Cheetham *et al*., 1992). The active site interactions described are summarised in Table 2 and shown in Figure 4.

Protein atom	**Subsite**	**Interaction**
Asn103OD1/ Asn103ND2	A	H-bond with O6
Asp101OD2	A	H-bond with N2 of the ligand *N*-acetyl side chain
Asp101OD2	B	H-bond with O6
Trp62	B	Hydrophobic interaction
Trp62NE	C	H-bond with O6
Trp63NE	C	H-bond with O3
Asn59N	C	H-bond with O7
Ala107O	C	H-bond with N2

Table 2 Active site interactions with the NAG_3 ligand

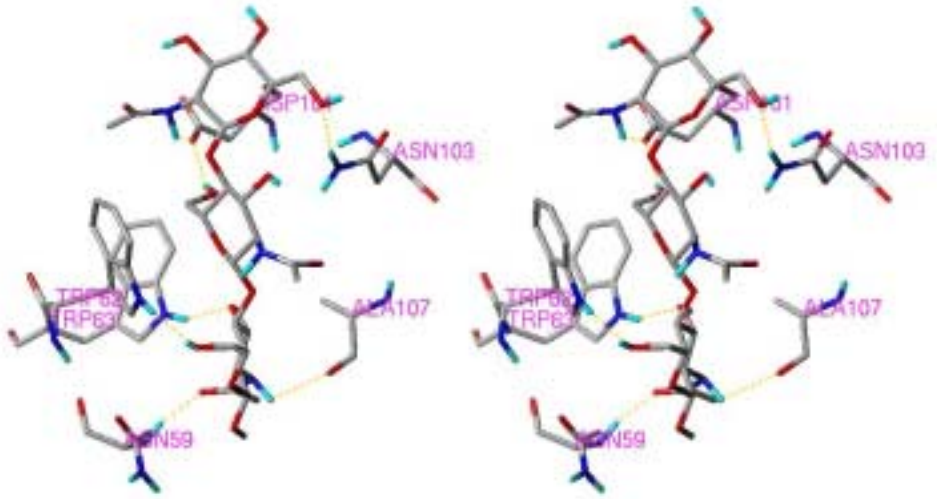


Figure 4 Stereo view. Residues in the HEWL active site involved in H-bonds (yellow dotted lines) to the NAG_3 ligand in PDB structure 1lzb.

Crystallographic studies of HEWL in several different crystal forms have shown that the catalytic residues Glu35 and Asp52 adopt almost identical conformations in all structures (Strynadka and James, 1996).

1.3.3 How comparable is lysozyme in the crystalline state to that in solution?

According to Strynadka and James, 1996, there have been studies verifying that lysozymes retain their activity within the constraints of the crystal lattice (Hadfield *et al.*, 1994; Howell *et al.*, 1992). There has also been a comparison of the three dimensional structure of HEWL as determined by X-ray crystallography and by NMR (Smith *et al.*, 1993) indicating a high degree of structural similarity.

2. Materials and Methods

2.1 Molecular modelling

The program package SYBYL (Tripos Inc.), versions 6.6 and 6.7 were used for all modelling. The modules SELECTOR and DIVERSE SOLUTIONS (DVS) (Pearlman *et al*., 1999) were used for compound selection. UNITY database tools, versions 4.1 and 4.2, were used to modify and explore the databases. SYBYL's Molecular Spreadsheet was used for data manipulations. The initial part of the work was carried out on an Silicon Graphics Indigo 2 workstation whereas some of the final and more computer-intensive searches and docking simulations were performed on an Silicon Graphics OCTANE workstation with two processors.

2.1.1 UNITY-databases

The ACD database consists of compounds from different commercial vendors. Some of these companies were regarded as more reliable and they were extracted from the ACD database. These databases were merged into one large database (called the mother database in this study and contains approximately 200 000 compounds); from which two UNITY-databases (one 2D and one 3D) were created. The 3D co-ordinates were generated using CONCORD (Pearlman). 2D-fingerprints and *macro*[5] screens were included in the databases to increase the speed of the screenings. For the same reason the large 3D-database was divided into 100 small databases with 2000 compounds each and the UNITY 2D database was divided into 5 smaller databases.

2.2 Virtual screening approaches

Different approaches have been used to search for new inhibitors to hen egg-white lysozyme within the ACD database. A flowchart of the methods is shown in Appendix 1. Two procedures (diversity selections) were based completely on diversity in metric descriptor spaces. Two approaches (similarity searches) were purely ligand-based. Two additional VS procedures (UNITY flexible search and FlexX docking) were mainly receptor-based. However some indirect information from the ligand was used in the latter (see below). In order to construct reliable and sensible receptor-based VS

[5]With *macro*, it is meant here, a feature found in a query. Some examples are the presence or absence of H-bond donors or acceptors and hydrophobic rings.

systems 11 crystal structures with a resolution better than 2.0 Å were superimposed and compared. All the different VS methods, with the exception of docking have been applied to the entire mother database. Docking has been combined with other methods in the following way. Diversity selections and ligand-based screenings were combined with docking and similarity searches in iterative protocols, and the output from the UNITY flexible search was submitted to docking.

2.3 Diversity selections

Two different programs, SELECTOR and DIVERSE SOLUTIONS (DVS) were used to select optimally diverse subsets from the 2D UNITY database *i.e.* the most dissimilar structures were selected. Using the *Optisim* algorithm in SELECTOR, one subset of 50 compounds was selected. SELECTOR uses 2D-fingerprints and the Tanimoto coefficient[6] to make the diversity selection. The Tanimoto coefficient for the *Optisim* selection was chosen to be 0.5, *i.e.* there were no pairs of selected structures with a Tanimoto Coefficient larger than 0.5. One subset of 1000 compounds was selected using the *distance-based* subset selection method in DVS. DIVERSE SOLUTIONS uses so-called B-cut values[7] as metric descriptors and DVS selects the descriptors that are most suitable for the database used. The B-cut descriptors based on H-bond donors, H-bond acceptors, Gasteiger charges (partial charges) and polarisability were used in these selections.

2.4 Ligand-based similarity searches

The same programs (but different algorithms) and even the same metrics used in the diversity selection were used to select compounds similar to the NAG ligands from the 2D UNITY database. The *dbsimilar* algorithm within SELECTOR uses the 2D-

[6] 2D-fingerprints describe the two dimensional structure of the compounds as a binary array. The array contains information about the presence and absence of different molecular fragments of particular lengths. Using 2D-fingerprints, the similarity between two structures can be measured with the Tanimoto coefficient. The binary arrays from the two compounds are compared. If they are exactly the same, which they can be even if they are isomers, the Tanimoto coefficient is 1.0. If the two arrays have no fragments in common the Tanimoto coefficient is 0.0. The Tanimoto coefficient assumes all values between 0.0 and 1.0.

[7] These descriptors are the lowest or highest eigenvalues of connectivity matrices, where the diagonal elements are calculated atomic properties such as hydrogen-bond posibilities, Gasteiger charges and polarisabilty. They include information about the molecular distribution of these properties. Distances are defined in the descriptor space in euclidean manner.

fingerprints to choose the structurally closest neighbours. The Tanimoto coefficient, defining the similarity radius for the neighbours, was adjusted between 0.76 to 0.89 to obtain approximately 50 compounds. For the *nearest-neighbours* selection algorithm in DVS the 100 closest neighbours were selected.

2.5 Comparison of the active sites of different crystal structures

There were almost 200 structures of hen egg-white lysozyme available in the Protein Data Bank (PDB), and 66 structures with a resolution better than 2.0 Å were superimposed (listed in Appendix 2). The three apo and the six liganded structures with the highest resolution were investigated more closely. Two additional high-resolution structures were chosen for comparison purposes. The structures were superimposed using the program BIOPOLYMER (Tripos Inc.) using all main-chain atoms for the superimposition. All amino-acid residues from the liganded structures that forms hydrogen bonds to the NAG ligands were tabulated and the relative frequency of the hydrogen bonding of each residue was calculated. Fast Conolly surfaces (Conolly, 1983) were calculated using MOLCAD (Tripos Inc.) for the active sites of these complexes. The surfaces were compared using the complex structure with the highest resolution (PDB code 1lzb) as the reference. For reasons that will be elaborated in the results and discussion section, both 1lzb and 1lks have been used for virtual screenings.

2.6 UNITY flexible search

A pharmacophore definition can be a description of the structural and functional features of a binding site. A pharmacophore was created and used as a query in a UNITY flexible search, to find compounds that could match its features. Some important H-bonds, a hydrophobic feature and a constraint surface were included in the final query. The complex structure with the highest resolution (PDB code 1lzb) was used in the definition of the UNITY query.

The most frequent hydrogen bonding residues were Asn59, Trp62 and Trp63. The side chain conformations of Asn59 and Trp63 are conserved among the structures. Therefore, two of the possible four H-bond donors belonging to these residues were included in the query. The crystal structure 1lzb has the highest

resolution (1.5 Å) of the NAG-containing structures and was therefore used as a template for the query. A hydrophobic feature was included based on the hydrophobic interaction between the apolar face of the saccharide ring in site B and the plane of the indole ring of Trp62. A constraint surface was also included in the query to limit the size and the shape of the compounds and to force them to fit into the active site. A fast Conolly surface created by MOLCAD (default probe radius of 1.4 Å) was used as a template for the query surface. Up to five different queries were constructed using different feature combinations and tested with a subset selected by the DVS distance-based algorithm (1000 compounds) from the mother database. The final query was such that 0.3% (3 molecules) of the subset were obtained as hits. The features of the final query are listed in Table 3 and depicted in Figure 5. This query was used in a UNITY flexible search. The flexible search protocol includes first a 2D-screen based on 2D-fingerprints followed by a Macro screen to reduce the number of structures to be tested by the flexible search. The default parameters were used in the search.

Feature Nr	**Feature type**	**Centre**	**Shape**	**Toler-ance**	**Associated Feature**
1	Protein donor	Asn59N	sphere	1.0 Å	5
2	Protein donor	Trp63NE	sphere	1.0 Å	6
3	Protein acceptor	Ala107O	sphere	1.0 Å	7
4	Hydrophobic	Centroid of NAG residue at site B	sphere	1.0 Å	
5	Ligand acceptor	~O7 site C	sphere	1.0 Å	1
6	Ligand acceptor	~O3 site C	sphere	1.0 Å	2
7	Ligand donor	~site C	cap	0.7 Å	3
8	Constraint Surface	NAG residue at site B	follows the active site shape	0.5 Å	

Table 3 Query definition. Features and constraints for the UNITY query, used in the UNITY flexible search. Protein donor number 1 is defined as a sphere with the radius of 1.0 Å around Asn59N on the protein and it is associated with the ligand acceptor number 5. UNITY searches for a ligand with an atom that is positioned within the sphere of the ligand acceptor number 5 and which has the possibility to make an H-bond to Asn59N. Protein donor number 2 and protein acceptor number 3 is used in the same way as protein number 1 and are associated with ligand acceptor number 6 and ligand donor number 7 respectively. The hydrophobic feature is placed were the centre of the sugar ring of NAG_3 in subsite B is positioned in the complex structure with PDB code 1lzb.

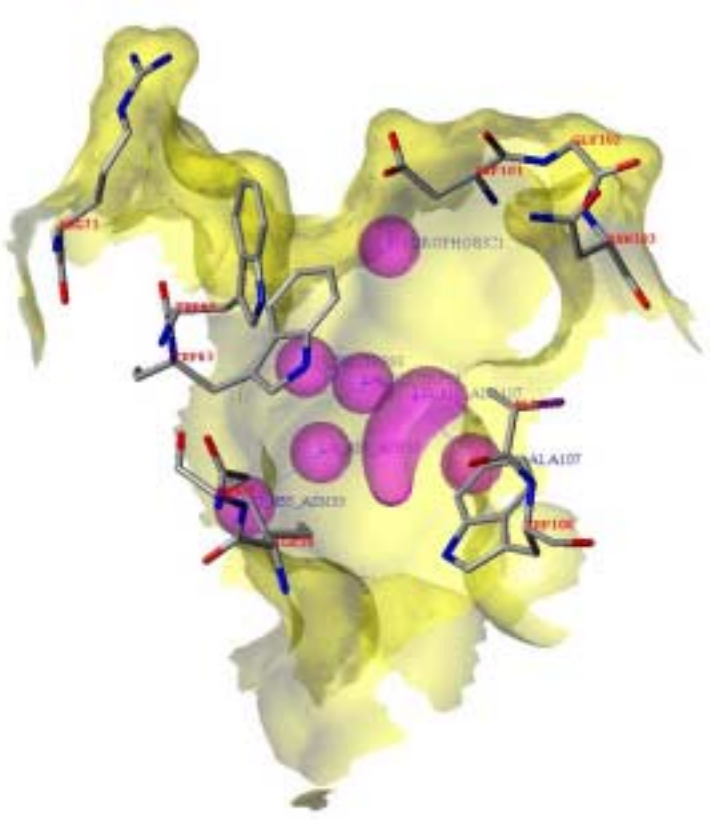


Figure 5 Query definition. This query was used to search for molecules that would match the properties of the active site of lysozyme. There are two hydrogen-bond donors on the protein (Trp63 and Asn59) and their associated features (ligand H-bond acceptors) are also shown. There is one hydrogen bond acceptor (Ala107) and its associated feature (ligand H-bond donor) is shown as a cap. There is also a hydrophobic feature (the top sphere in pink) and a surface (yellow).

2.7 FlexX docking simulations

FlexX is a fast docking method that allows torsional flexibility in the ligands while keeping the receptor fixed. It places ligands into the active site starting with one fragment, and subsequently builds the whole molecule into the active site fragment by fragment. FlexX uses formal charges, which were turned on during the docking simulations. All the relevant receptor information necessary for the docking simulations is stored in the receptor definition file (rd file or rdf). Two different structures have been used as a base for docking simulations to take into consideration alternative conformations (see the Results and discussion section). The corresponding PDB codes are 1lzb and 1lks[8]. Defaults have been used when creating the rd files and no special customisations have been done. Essentially then, the residues belonging to the active site file in the rd file determine the docking method uniquely. The active

site residues for the 1lzb structure were chosen to be all those residues with at least one atom within 4.0 Å from NAG_3 ligand[9]. These residues are D52, L56, N57, I58, N59, W62, W63, R73, L75, I98, D101, G102, N103, A107 and W108. The residues in the active site file corresponding to 1lks are E35, N46, D48, S50, T51, D52, L56, Q57, I58, N59, S60, R61, W62, W63, I98, N103, N106, A107, W108 and V109.

2.8 Diversity and similarity searches combined with FlexX docking

Two iterative protocols have been developed depending on the programs used. In Protocol 1a, algorithms from the SELECTOR program have been employed whereas in Protocol 1b, the corresponding program from DIVERSE SOLUTIONS (DVS) have been used.

CONCORD was used to generate the 3D co-ordinates of the molecules in the two subsets selected by diversity selection. They were minimised with the MMFF94s force field (Mastryukov *et al*., 1993) and docked using FlexX. The 50 molecules selected by means of SELECTOR's *Optisim* algorithm were docked into the active sites of both 1lzb and 1lks whereas the 1000 molecules selected with DVS *distance-based* algorithm were docked only into 1lzb since this is the liganded structure. The FlexX energy was used to rank the molecules. The 10 molecules with highest rank in the SELECTOR subset of 50 (Protocol 1a) and the 50 best in the DVS subset of 1000 (Protocol 1b) were used as starting points for similarity searches using the corresponding algorithms (SELECTOR's *dbsimilar* in Protocol 1a and *nearest-neighbours* DVS in Protocol 1b). The closest neighbours (50 in Protocol 1a and 1000 in Protocol 1b) were docked again and the procedures continued in an iterative fashion, each time taking the 10 (Protocol 1a) and 50 (Protocol 1b) best ranked compounds from (*all*) the previous dockings as target for similarity selections. After docking the originally diversity-based selected molecules, 4 rounds of similarity selection and docking were performed. Before the molecules from the DVS selection were docked, all compounds with more than 20 rotatable bonds were filtered out to reduce the entropic effects due to binding (because FlexX cannot predict these).

---

[8] FlexX automatically ignores hetero-atoms. Therefore, it was not necessary to remove the ligand or solvent molecules from the PDB files.

[9] Using this receptor definition file it was possible to reproduce the conformation of the NAG_3 complex by means of docking simulation.

2.9 Ligand-based searches combined with FlexX docking

As in the diversity–based methods, two iterative protocols have been developed: SELECTOR has been used in Protocol 2a and DVS in Protocol 2b.

The 3D co-ordinates for the 50 compounds selected by the similarity search using SELECTOR and the 100 compounds selected by DVS were generated using CONCORD. The molecules were minimised using the MMFF94s force field. FlexX was used to dock the compounds into the active site. The molecules selected by means of SELECTOR's *dbsimilar* algorithm were docked into both 1lzb and 1lks active sites (Protocol 2a) whereas the 100 molecules selected with DVS *nearest neighbour* algorithm were docked only into 1lzb (Protocol 2b). The docking energies generated by FlexX were used as ranking scores and the 10 best compounds were used as the starting compounds for a second round of similarity searches, each time taking the best ranked compounds from (*all*) the previous dockings as target for similarity selections. After docking the originally ligand-based selected molecules, 4 rounds of similarity selection and docking were performed.

2.10 FlexX docking of the UNITY flexible search results

The hits from the UNITY flexible search were minimised using the MMFF94s force field and docked into the active site of 1lzb using FlexX to further scrutinise these results.

3. Results and discussion

3.1 Diversity selections

Two diverse compound subsets were obtained, one of 50 compounds from the *Optisim* selection and one of 979 compounds from the *distance-based* DVS selection (see attached files optisim.sdf and selected_distance.sdf if available). Since these two subsets were selected independently from the protein target and the known ligands, they are general and can be used to search for ligands to any target. *Optisim* uses a random seed molecule as the starting point in the selection procedure and therefore the result will be different in a second selection, although the diversity will be roughly the same. In the *distance-based* DVS selection compounds are selected randomly, and all molecules within a certain distance from a selected compound are eliminated before the next compound is selected. The order of this algorithm is linear with respect both to the size of the starting set and to the subset size, and is therefore far faster than most other methods and is more suitable for large databases.

The diversity of the two subsets is visualised in the property space of the B-cut metrics based on H-bond donors and acceptors (Figure 6) and on Gasteiger charges and polarisability (Figure 7). The compounds in the diverse subsets are well scattered in the property space defined by these metrics, especially in the Gasteiger charges/polarisability metrics that describe the distribution of partial charges and hydrophobicity on molecules. The mean Tanimoto coefficient was also calculated for the two subsets (Table 4). The mean Tanimoto coefficient (to the closest neighbour) is higher for the *distance-based* subset (0.51) than for the *Optisim* subset (0.26). This is almost certainly due to the much larger number of compounds in the *distance-based* subset.

Subset	**Nr of compounds in the subset**	**Mean Tanimoto (to nearest neighbour)**	**std dev (mean Tanimoto)**
DVS *distance-based*	979	0.51	0.15
SELECTOR O*ptisim*	50	0.26	0.06
DVS *nearest neighbours*	100	0.86	0.23
SELECTOR *dbsimilar*	48	0.95	0.19
Protocol 1b (best 50 after docking 5)	50	0.78	0.17
UNITY query	994	0.79	0.16
UNITY query (best 50 after one docking)	50	0.69	0.19

Table 4 Comparison of the diversity of the different subsets, the comparison is based on 2D-fingerprints. The mean Tanimoto coefficient is the average distance between every compound and its closest neighbour within the subset. The DVS distance-based subset is diverse measured with the 2D-fingerprint descriptors.

Both the *Optisim* selection and the *distance-based* DVS selection were fast (minutes), however the *Optisim* method used here was not exhaustive, an exhaustive *Optisim* selection would have taken considerably longer time to perform for this large database. DVS is therefore more suitable for selections from large databases (more than 50000 compounds) whereas *Optisim* works very well for smaller databases (less than 50000 compounds).

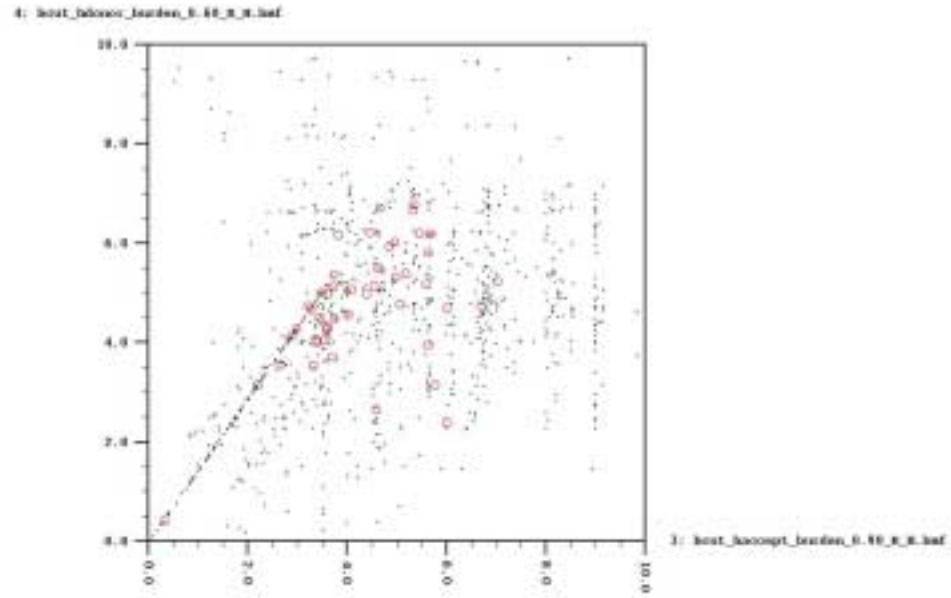


Figure 6 The spread of the compounds from the *distance-based* and the *Optisim* subsets in the B-cut metrics based on H-bond acceptors and H-bond donors. The *distance-based* subset is shown in black dots and the *Optisim* subset in red circles. Both subsets are diverse but the smaller *Optisim* subset is distributed around the centre. The line of dots from the bottom left corner is most probably due to molecules where all H-bond donors are also H-bond acceptors.

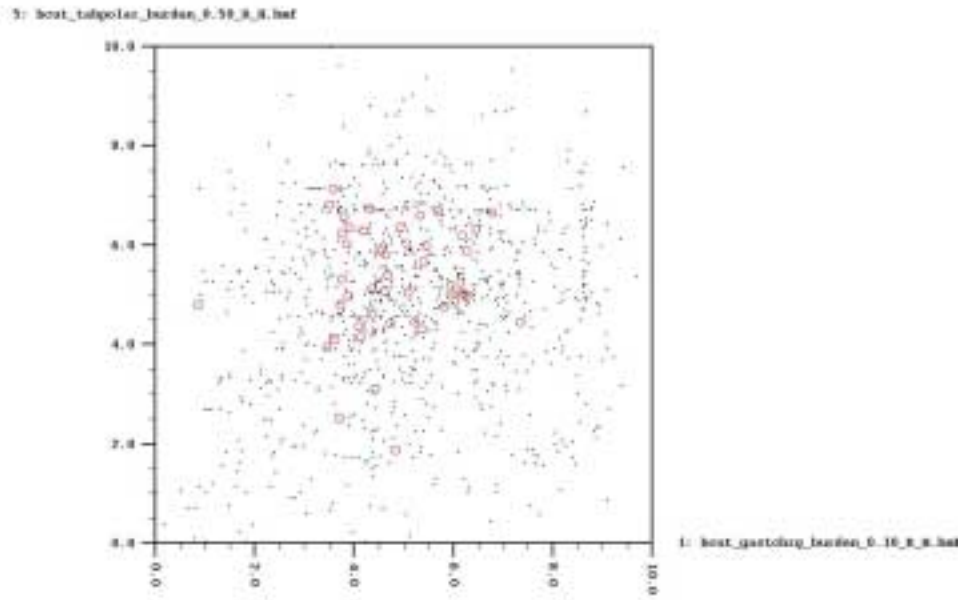


Figure 7 The spread of the compounds from the *distance-based* and the *Optisim* subsets in the B-cut metrics Gasteiger charges and polarisability. The *distance-based* subset is shown in black dots and the *Optisim* subset in red circles. Both these subsets are diverse with a slight concentration around the centre for the *distance-based* subset and with no compounds around the edges for the *Optisim* subset.

3.2 Ligand-based similarity searches

Two subsets of compounds similar to the NAG ligands were obtained, one subset of 48 compounds from the SELECTOR search (*dbsimilar*) and one subset of 100 compounds from the DVS selection (*nearest-neighbours*) (see attached files dbsim1.sdf and nearest1.sdf if available). These two subsets were selected using two different methods based on different types of descriptors. However, the findings from the selections were very similar. 35 compounds selected by *dbsimilar* were also selected by *nearest-neighbours.* All molecules from the *dbsimilar* selection contained sugars or sugar-like molecules, whereas approximately 75 percent of the structures selected by DVS contained sugar rings and 25 percent were other types of compounds. This was expected since the *dbsimilar* method selects only structurally similar molecules while DVS uses *property* descriptors. The *nearest-neighbour* selection

seems to cover the findings from the *dbsimilar* selection and finds also molecules with similar properties but dissimilar structure.

The spread of the molecules in the property space defined by the B-cut metrics based on Gasteiger charges and polarisability is visualised in Figure 8, the chemistry space for the distance-based diversity selection is also shown for comparison. The two subsets from the different selection methods are clustered in the same area. The mean Tanimoto coefficient for the *dbsimilar* subset was defined from the beginning, the molecules from the *nearest-neighbours* selection are more dissimilar but that is probably because it is a larger subset (Table 4).

DVS is more suitable for similarity searches in large databases. The *nearest-neighbour* selection took only a few seconds for the whole mother database, while the *dbsimilar* selection became extremely slow for databases larger than 50 000 compounds, when the mother database was divided into five smaller databases the *dbsimilar* selection took a few minutes.

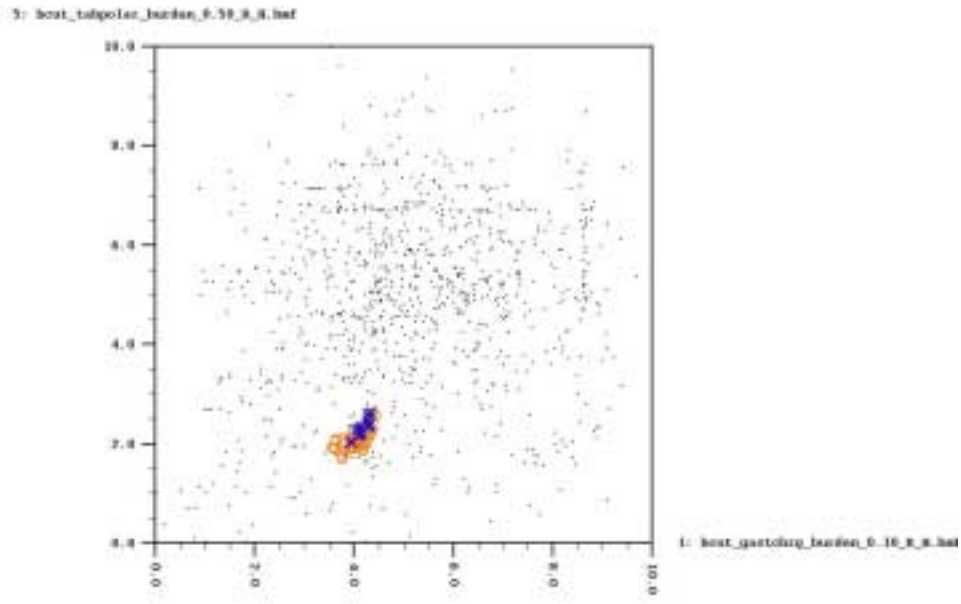


Figure 8 The spread of the compounds from the *nearest-neighbours* and the *dbsimilar* subsets, selected using the NAG ligands as starting point, in the B-cut metrics based on Gasteiger charges and polarisability. The *distance-based* subset (black dots) is shown for comparison together with the *nearest-neighbours* selection (orange circles) and the *dbsimilar* subset (blue X's). The two subsets from the different selection methods are clustered in the same area.

3.3 Comparison of the active sites of different crystal structures

Among the more closely investigated crystal structures it was possible to define two different groups of molecules (Table 5). One large group contains 9 structures both with and without ligands, the other group consisted of only two structures (1lks and 4lzt, both without ligands). The large differences were the position of Trp62, which pointed straight into the active site in group 2 and the peptide flip between Asn103 and Gly104. In the absence of a ligand, the side chain of Trp62 has the possibility of adopting alternative conformations (Figure 3) depending most probably on the type of crystal contacts. This is possible since the packing of the group 2 structures are different from the space groups of group 1 structures (Table 5). It is also possible to see the movement of Trp62 between the apo protein and the liganded structures in Figure 3. The findings from the comparison of the H-bonds are shown in Table 6.

PDB-code	Resolution (Å)	Ligand	Rmsd (Å)	Space group	Group
1lzb	1.50	NAG_3	0.0	$P4_32_12$	1
1lza	1.60	-	0.3	$P4_32_12$	1
1lzc	1.80	NAG_4	0.4	$P4_32_12$	1
1hew	1.75	NAG_3	0.7	$P4_32_12$	1
193l	1.33	-	0.5	$P4_32_12$	1
1aki	1.50	-	1.1	$P2_12_12_1$	1
1at5	1.80	NAG_3	0.4	$P4_32_12$	1
1at6	1.80	NAG_3	0.8	$P4_32_12$	1
1uih	1.75	NAG_3	0.5	$P4_32_12$	1
4lzt	0.95	-	1.3	P1	2
1lks	1.10	-	1.4	$P2_1$	2

Table 5 Comparison of different crystal structures from the PDB-database. The root-mean-square distance (r.m.s.d.) is given in relation to 1lzb and is based on the main-chain atoms. The structures are grouped according to the position of the residues.

PDB code	Rmsd to 1lzb (Å)	Ligand	Asn46	Asn59 N	Trp62 NE	Trp63 NE	Asp101 OD2	Asn103 OD2/ ND2	Ala107 O
1lzb	reference	NAG_3	1	1	1	1	1	1	1
1hew	0.7	NAG_3	0	1	1	1	1	1	1
1lzc	0.4	NAG_4	1	1	1	1	1	0	1
1uib	-	NAG_3	0	1	1	1	1	0	0
1at5	0.4	NAG_3	0	1	1	1	0	0	1
1uih	0.5	NAG_3	0	1	1	1	1	0	1
1at6	0.8	NAG_3	0	1	1	1	0	0	0
Relative frequency			0.29	1.0	1.0	1.0	0.71	0.29	0.71

Table 6 The relative frequency of H-bonding among amino-acid residues to the NAG ligands in some crystal structures from the PDB.

3.4 UNITY flexible search

A total of 994 molecules were found to match the final UNITY query (see attached file UNITY_query.sdf if available). Many of the molecules were long and linear, almost all contained one or several phenyl groups and there were only a few sugar-like compounds. The spread of these molecules in the chemical property space defined by the B-cut metrics based on H-bond donors, H-bond acceptors, Gasteiger charges and polarisability is shown in Figure 9 and Figure 10. The *distance-based* subset from the DVS diversity selection is shown for comparison. H-bond donor and H-bond acceptor properties are included in the query and therefore the B-cut values based on these are not unbiased descriptors for this subset of hits. In the chemistry space defined by the molecular distribution of Gasteiger charges and polarisability the hits are clustered in a small area. This indicates that these descriptors are relevant for the receptor since they are not as biased by the query as the B-cuts based on H-bond donors and acceptors. This also indicates the existence of an “island of activity” from which the majority of the hits originate.

This search was a relatively CPU-time consuming process, it took 13 days to search through the 3D UNITY database of approximately 200 000 compounds using the Silicon Graphics OCTANE workstation. There are faster search algorithms within UNITY but the flexible search tries a larger number of conformations (within a time limit) of the compounds relevant for the query, and that gives more reliable answers

than trying just a few different conformations. Some compounds “timed out” and have therefore not been included in the results.

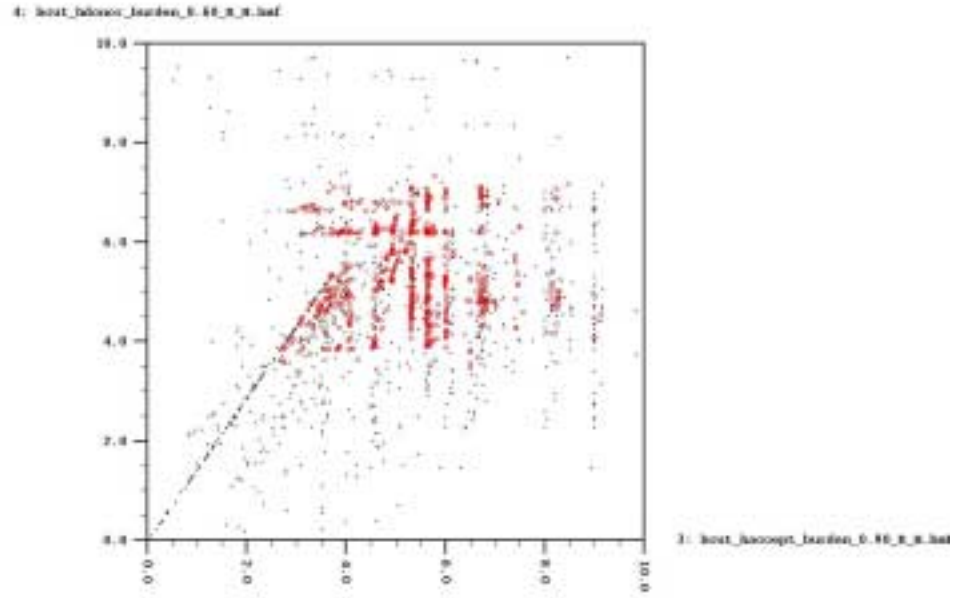


Figure 9 The spread of the compounds from the UNITY flexible search in the B-cut metrics based on H-bond acceptors and H-bond donors. The *distance-based* subset (black dots) is shown for comparison and the result from the UNITY query is shown in red circles. These descriptors are somewhat biased by the query since the H-bond donors and acceptors are included in the query.

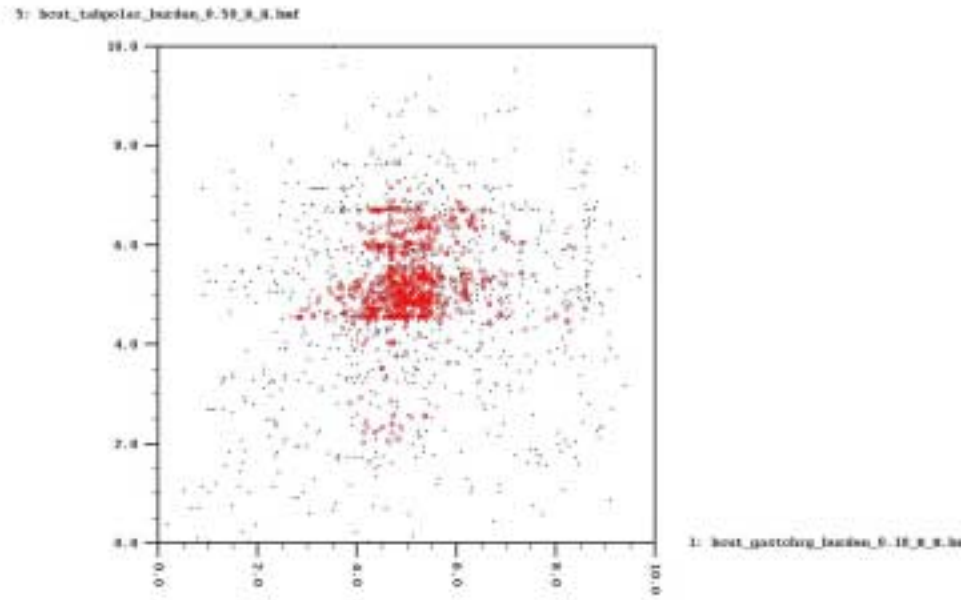


Figure 10 The spread of the compounds from the UNITY flexible search in the B-cut metrics based on Gasteiger charges and polarisability. The distance-based subset (black dots) is shown for comparison. The result from the UNITY query (red circles) is clustered and that indicates a relevance for these descriptors, which are not as biased by the query as the B-cuts based on H-bond donors and acceptors.

3.5 Diversity and similarity searches combined with FlexX docking

3.5.1 Iterative protocol 1a: using SELECTOR's Optisim and dbsimilar

The mean energy of the 10 best hits (see attached file opti5_1lzb_10.sdf if available) from the last docking (into PDB structure 1lzb) was -30±1 kJ/mol, the minimum energy was –31±1 kJ/mol and the maximum was –29±1 kJ/mol. In Table 7 the mean energies from the 10 best compounds of all five dockings are shown and it is possible to see how the energy decreased after each round, faster in the beginning and slower in the end (Figure 11). The H-bonds to Asp52 and Trp63 were the most frequently found interactions. A few H-bonds were formed to Trp62 and Asn59, but none to Ala107, in variance with what is shown in Table 6.

Figure 11 Mean energies for the 10 best (50 best for Protocol 1b, green triangles) compounds after each cycle. Protocol 1a departing from SELECTOR's *Optisim* subset shown in dark blue diamonds (docked into PDB structure 1lzb) and in pink squares (docked into PDB structure 1lks), Protocol 1b departing from DVS *distance-based* subset shown in green triangles. Protocol 2a using SELECTOR's *dbsimilar* shown in orange asterisks (docked into PDB structure 1lzb) and in blue circles (docked into PDB structure 1lks) and Protocol 2b DVS *nearest-neighbours* shown in light blue X's.

Round	**Method**	**Tanimoto**	**Mean Energy (kJ/mol)**	**Std dev Energy (kJ/mol)**
0	*Optisim*	0.50	-19±3	3.25
1	*dbsimilar*	0.80	-27±2	2.10
2	*dbsimilar*	0.75	-28±1	1.46
3	*dbsimilar*	0.68	-29±1	1.20
4	*dbsimilar*	0.63	-30±1	0.90

Table 7 Tanimoto coefficient, mean energies and standard deviations of the 10 best compounds from FlexX docking into PDB structure 1lzb of subsets selected using SELECTOR's algorithms: *Optisim* diversity selection and *dbsimilar* (Protocol 1a). The Tanimoto coefficient in the first round was used as the threshold value for *dissimilarity, i.e.* no pairs of structures had a Tanimoto coefficient larger than 0.50. In the other rounds the Tanimoto coefficient was used for *similarity, i. e.* all compounds had a Tanimoto coefficient with one of the starting compounds which was larger than the value indicated.

The 10 best compounds (see attached file opti5_1lks_10.sdf if available) from the last docking (into PDB structure 1lks) had a mean energy of -30±1 kJ/mol, a minimum of –32±1 kJ/mol and a maximum of –29±1 kJ/mol. These molecules had mostly H-

bonds to Asp52, Gln57 and Val109, only very few H-bonds were made to the residues mostly used in the interactions between HEWL and NAG_3 (Asn59, Trp63 and Ala107). The mean energies from the 10 best compounds from all five dockings are listed in Table 8 and shown in Figure 11.

Round	**Method**	**Tanimoto**	**Mean Energy (kJ/mol)**	**Std dev Energy (kJ/mol)**
0	*Optisim*	0.50	-21±5	4.63
1	*dbsimilar*	0.75	-28±2	2.24
2	*dbsimilar*	0.78	-28±2	1.85
3	*dbsimilar*	0.75	-30±1	1.25
4	*dbsimilar*	0.66	-30±1	1.08

Table 8 Tanimoto coefficient, mean energies and standard deviations of the 10 best compounds from FlexX docking into PDB-structure 1lks of subsets selected using SELECTOR's algorithms: *Optisim* diversity selection and *dbsimilar* (Protocol 1a). The Tanimoto coefficient in the first round was used as the threshold value for *dissimilarity, i.e.* no pairs of structures had a Tanimoto coefficient larger than 0.50. In the other rounds the Tanimoto coefficient was used for *similarity, i. e.* all compounds had a Tanimoto coefficient with one of the starting compounds which was larger than the value indicated.

The 10 best structures from each of these two iterative protocols were very similar, even though the starting set from the *Optisim* selection was diverse (Figure 6 and Figure 7). This may be due in part to the way the similarity searches were carried out. Ten molecules were used as input to the similarity search, *dbsimilar*, but only between two and four of these 10 had neighbours within the specified similarity radius (Tanimoto coefficient). Neighbours were therefore only selected to a minor subset of the input structures and the diversity of the subset was lost relatively fast. Another reason may be the small starting subset (50 molecules), which cannot cover as much of the total property space of the mother database as the larger subset (1000 molecules) selected by DVS. The position of these results in the chemistry space is shown in Figure 12. Both these subsets are clustered in a very small area.

The computer time consumed by these iterative processes were as follows: similarity selections, minutes; minimisations, two to three hours; and docking, six to nine hours, all using the Silicon Graphics Indigo 2 workstation.

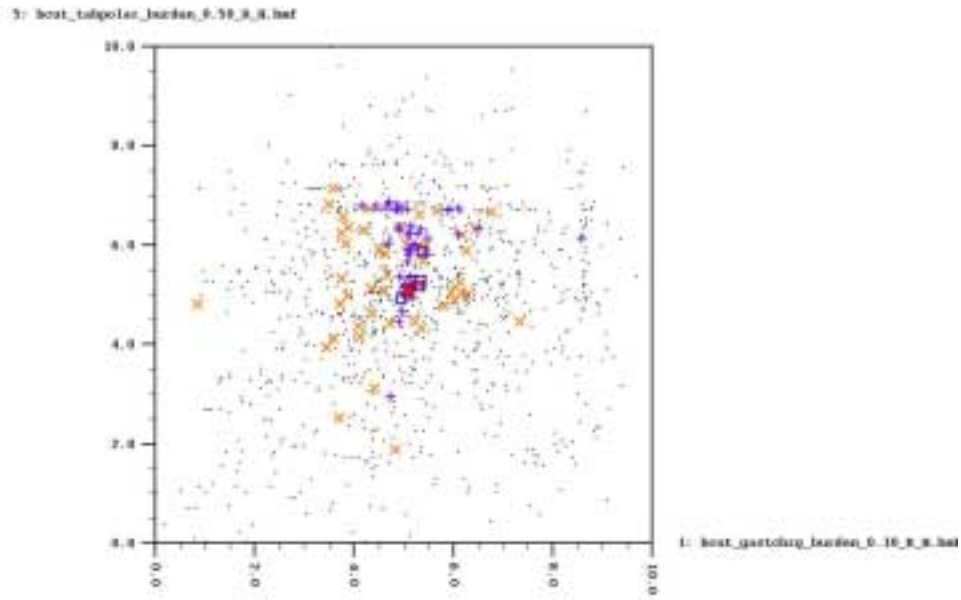


Figure 12 The spread of the compounds from Iterative Protocols 1a and 1b in the B-cut metrics based on Gasteiger charges and polarisability. The *distance-based* starting subset (black dots) and the *Optisim* starting subset (orange X's) are shown for comparison. The resulting compounds from Protocol 1a are clustered in a very small area (docked into PDB structures 1lzb, red circles, and 1lks, blue squares) this area is part of the "island" where the compounds from Protocol 1b (purple +'s) are clustered.

3.5.2 Iterative protocol 1b: using DVS distance-based and nearest-neighbours

The mean energies from the 50 highest ranked compounds from each round of selections are presented in Table 9 and it is possible to see how the predicted affinity for the ligands increased after every iterative cycle (Figure 11). The minimum energy from the 50 best compounds of the last docking was –38.00 kJ/mol and the maximum was -32.50 kJ/mol.

Round	Method	Mean energy (kJ/mol)	std dev (Energy kJ/mol)
0	*Distance-based*	-25±3	2.60
1	*Nearest-nghbs*	-30±2	2.23
2	*Nearest-nghbs*	-32±2	1.65
3	*Nearest-nghbs*	-34±2	1.68
4	*Nearest-nghbs*	-35±2	1.51

Table 9 Mean energies and standard deviations of the 50 best compounds from FlexX docking of subsets selected using DVS algorithms: *distance-based* diversity selection and *nearest-neighbour* selection (Protocol 1b).

By superimposing all the 50 best ranked structures from the docking, some common features were shown (Figure 13). There is often (~30 out of 50) a phenyl ring close to the position of the saccharide ring in subsite B. This is in agreement with the hydrophobic stacking interaction described (Cheetham *et al*., 1992) between the Trp62 and the apolar face of the sugar residue. However, the phenyl rings were almost always positioned deeper into the active site cleft than the saccharide ring, probably because they are less bulky. In almost all structures (~40 out of 50), there is an oxygen atom accepting a proton from the amide nitrogen of Asn59, very close to the position of the carbonyl oxygen of the NAG-ligand *N*-acetyl group in site C (Figure 13). This is in agreement with the interaction that Asn59N makes with the ligand in the complex structures (Table 6). In a similar way it was found that often (~40 out of 50) a nitrogen atom is close to the position of the nitrogen of the ligand *N*-acetyl group in the liganded structures, hydrogen binding to Ala107O as in the complex (Table 6).

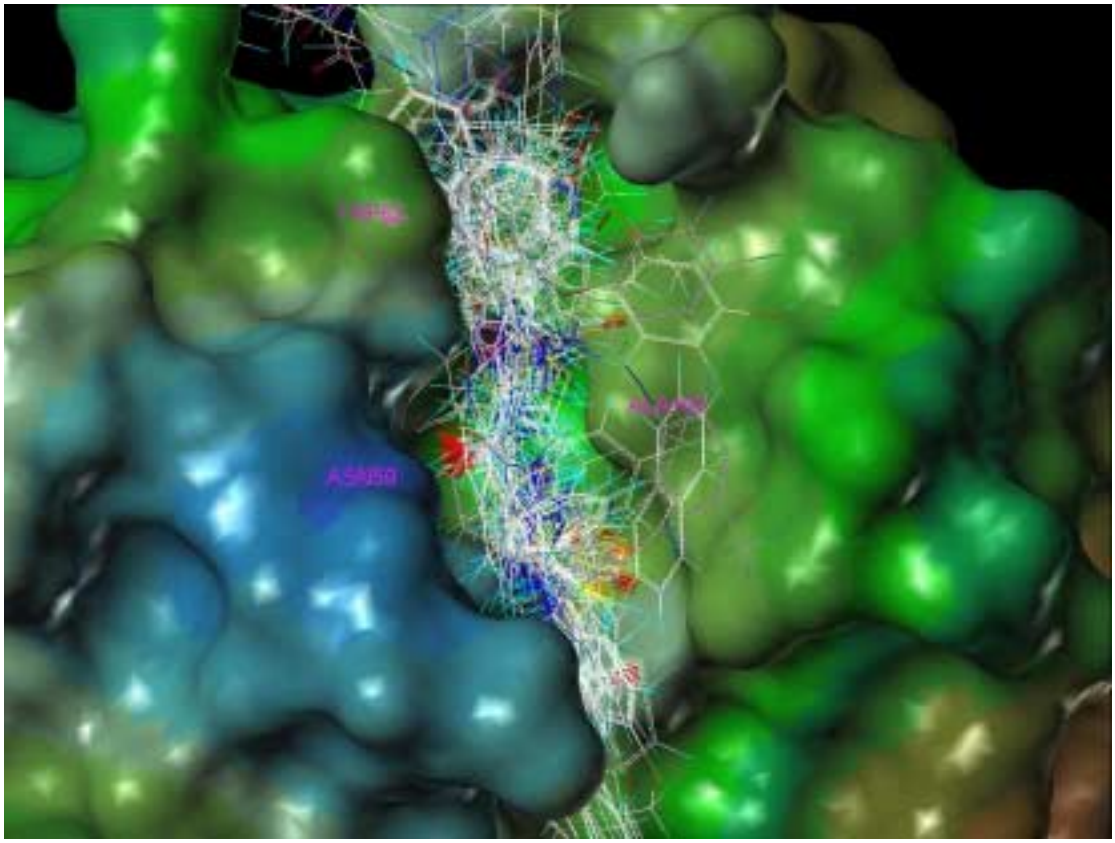


Figure 13 The 50 highest ranked compounds from the last docking of the hits from Protocol 1b. The large number of ring structures close to Trp62 is visible. A large number of carbonyl oxygens are clustered near Asn59. The nitrogens near Ala107 are not as tightly clustered because the direction of the H-bond from Ala107 is more variable. Two additional H-bonding residues were found Gln57 and Asp52, and a cluster of nitrogens is visible close to these residues, below the oxygen cluster close to Asn59.

In many of the structures (~35 out of 50) there is either a nitrogen or an oxygen atom accepting a hydrogen-bond from Trp63, close to the position of the carbonyl oxygen (O3) of the NAG-ligand, similar to the interaction between Trp63NE1 and the ligand (Table 6). However, Trp62NE is rarely involved in H-bonds, possibly because the hits are positioned deeper into the active site than the larger NAG molecules whereas Trp62 is situated in the outer parts of the active site. Two additional H-bond donor residues were found namely: Asp52 and Gln57. Asp 52 is one of the catalytic residues. A large number of hits (~40 out of 50) had one or two nitrogens close to one or both of these residues.

During this iterative process, molecules that fitted very well into the active site were found. There were no sugar-like structures in these results, but many long structures containing aromatic rings were filling the whole active site (Figure 14) that is normally filled by a linear hexa-saccharide substrate. The results from the last

cycles have been compared with the starting sets in the B-cut metrics based on Gasteiger charges and polarisability Figure 12. The compounds are clustered in a narrow area by the B-cut metric Gasteiger charges and in a quite narrow area by the polarisability metric. This iterative procedure apparently allows optimisation of the compounds coming from the original diversity selection.

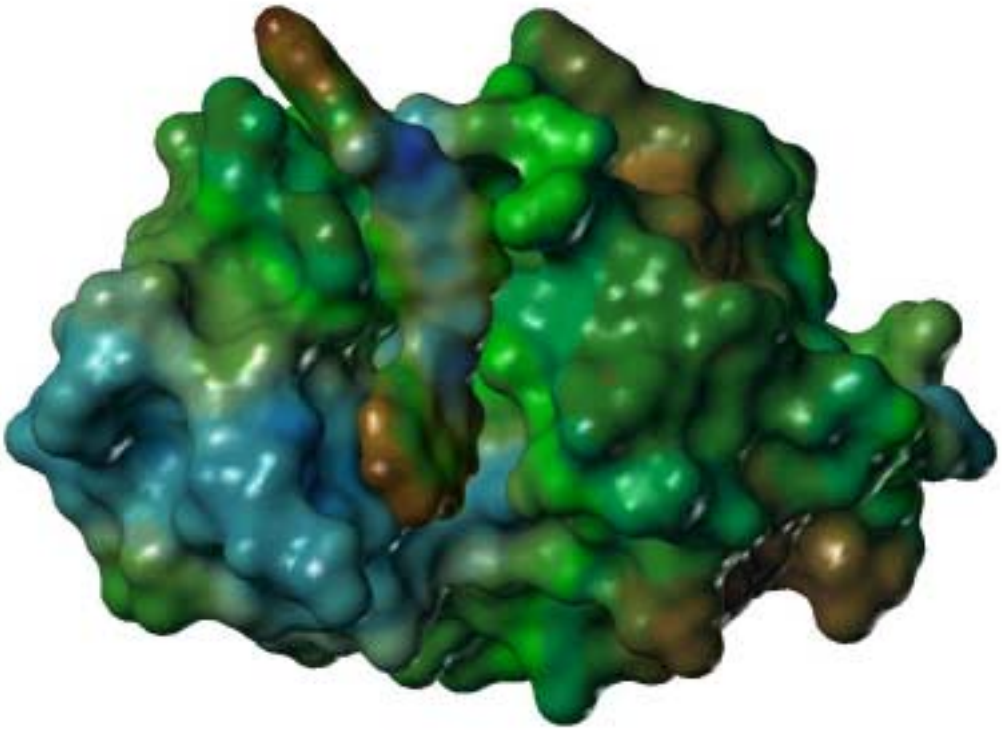


Figure 14 Fast Conolly surfaces of HEWL and one of the hits from virtual screenings found by Protocol 1b.

3.5.3 Comparison of the results from Protocol 1a and 1b

There is an interesting result obtained from these two independent iterative protocols. The resulting compounds from both protocols are clustered in a small and common area in chemistry space (Figure 12). This area is consistent with the "island of activity" found from the UNITY search results (see section 3.4).

The similarity selections were very fast (seconds to minutes), the minimisation processes took two to three hours and the docking took about 24 hours for each cycle (1000 compounds, using the Silicon Graphics OCTANE workstation).

3.6 Ligand-based searches combined with FlexX docking

3.6.1 Iterative protocol 2a: using SELECTOR's dbsimilar

The results in terms of energies of the best 10 compounds (see attached files lig5_1lzb_10.sdf and lig5_1lks_10.sdf if available) after each cycle are shown in (Table 10, Table 11 and Figure 11). The structures clustered almost exclusively around the subsites C and D. All structures contained one, two or three sugar rings and very few contained phenyl rings. Three structures contained aromatic rings close to the hydrophobic area around Ile98. These compounds were only selected based on structural similarity to the NAG ligands, and this was clearly reflected in the results.

There were no structures occupying site B in the subset selected using the PDB structure 1lks, because Trp62 occupies site B in this structure. However, there were also very few structures from the selection using 1lzb that occupied site B.

Round	**Method**	**Tanimoto**	**Mean Energy (kJ/mol)**	**Std dev Energy (kJ/mol)**
0	*dbsimilar*	0.85	-22±1	1.41
1	*dbsimilar*	0.89	-23±2	1.87
2	*dbsimilar*	0.80	-24±2	1.67
3	*dbsimilar*	0.85	-24±2	1.55
4	*dbsimilar*	0.76	-25±2	1.62

Table 10 Tanimoto coefficient, mean energies and standard deviations of the 10 best compounds from FlexX docking into PDB-structure 1lzb of subsets selected using SELECTOR's *dbsimilar* algorithm (Protocol 2a).

Round	**Method**	**Tanimoto**	**Mean Energy (kJ/mol)**	**std dev (Energy kJ/mol)**
0	*dbsimilar*	0.85	-25±2	2.26
1	*dbsimilar*	0.89	-26±2	2.35
2	*dbsimilar*	0.82	-26±2	2.27
3	*dbsimilar*	0.81	-26±2	1.81
4	*dbsimilar*	0.79	-27±2	2.32

Table 11 Tanimoto coefficient, mean energies and standard deviations of the 10 best compounds from FlexX docking into PDB-structure 1lks of subsets selected using SELECTOR's *dbsimilar* algorithm (Protocol 2a).

3.6.2 Iterative protocol 2b: using DVS's nearest-neighbours

The results in terms of energies of the best 10 compounds (see attached files dvs_ligand_nghbs_10.sdf if available) after each cycle are shown in Table 12 and

Figure 11. Hydrogen bonds were more frequent to Asp101 and Asn103 and there were no hydrogen bonds at all to Trp62. The H-bonds included in the UNITY query (Ala107, Trp63 and Asn59) were found moderately often. Many of the structures in this selection were small peptides. None of the 10 best-ranked compounds contained sugar rings even though the structures used as the starting point were only sugars. This clearly shows that this method selects compounds with similar physico-chemical features, rather than similar structures.

Round	**Method**	**Mean Energy (kJ/mol)**	**std dev (Energy kJ/mol)**
0	*Nearest-nghbs*	-21±4	3.55
1	*Nearest-nghbs*	-25±2	2.27
2	*Nearest-nghbs*	-25±2	2.13
3	*Nearest-nghbs*	-25±2	2.05
4	*Nearest-nghbs*	-25±2	2.05

Table 12 Mean energies and standard deviations of the 10 best compounds from FlexX docking of subsets selected using DVS algorithm: *nearest-neighbour* selection (Protocol 2b).

3.6.3 Comparison of the results from Protocol 2a and 2b

The 10 best compounds from the last rounds of similarity selections and docking, from all three methods, were superimposed and it was possible to see a pattern of H-bonds. Ala107O and Asn59N were the most frequently used hydrogen bonding atoms. This is similar to the interaction that Asn59N and Ala107O make with the ligand in the complex structures (Table 6). Other atoms also often involved in the H-bonds are Asp52OD1, Trp63NE, Gln57OD1 and, less often, Trp62NE.

The positions of these three subsets of ten compounds each, in the chemical property space are shown in Figure 15. The resulting compounds from both protocols are clustered in the same small area in chemistry space. This is the same area as the starting compounds that are similar to the NAG ligands. It is tempting to conclude then, that this area defines the natural “island of activity” where the substrate belongs.

The computer time consumed by each cycle of these iterative processes were; similarity selections, minutes; minimisations, two to three hours; and docking, 10 to 12 hours, all using the Silicon Graphics Indigo 2 workstation.

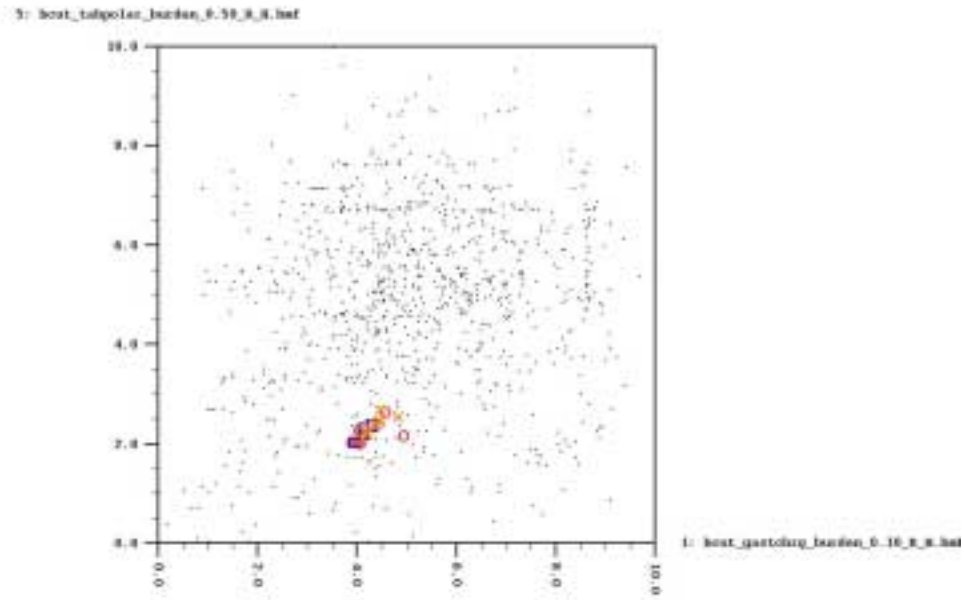


Figure 15 The spread of the compounds from Protocol 2a and 2b in the B-cut metrics based on Gasteiger charges and polarisability. The *distance-based* subset (black dots) is shown for comparison. The result after four cycles of *dbsimilar* selections and five dockings into PDB structures (1lzb, red circles and 1lks, blue squares) and the result after five cycles with DVS (orange X's) are all clustered in the same area. This is the same area where the initial DVS and SELECTOR similarity subsets, selected from the NAG ligands, were found.

3.7 FlexX docking of the UNITY flexible search results

The docking energies for the 50 best structures (see attached file unity_query_best_docked_50.sdf if available) varied between –39.10 kJ/mol and -26.50 kJ/mol, the mean energy was –29±3. Those relatively low predicted binding energies from FlexX validate the UNITY flexible search results. The iterative process starting from the DVS diversity selection above also found the structure with the best energy in this docking.

3.8 Final discussion

The increasing predicted affinity between HEWL and the ligands is clearly visible in Figure 11 especially for Protocol 1b, where 1000 new compounds were selected each cycle, and which shows a decrease in the mean docking energy of almost 10 kJ/mol.

The interactions found to be important by studying the complex crystal structures and by trying some different UNITY queries were also verified by the results from the molecules found by Protocol 1b. The three hydrogen-bonding atoms, chosen to be part of the UNITY query (Asn59, Trp63 and Ala107), were also frequently involved in H-bonds of the ligands found Protocol 1b. This was also in accordance with the H-bonding residues involved in the NAG interactions, except from Trp62 (Table 6). Trp62 was left out of the UNITY query despite the fact that it occurs in all NAG complexes (Table 6), and that was found to be in accordance with the result from Protocol 1b. The hydrophobic interaction of the NAG ligand in subsite B was used as a template for the hydrophobic interaction in the UNITY query. The importance of this interaction was also verified by the large number of hydrophobic ring-structures in subsite B in the result from Protocol 1b.

Both the diverse subsets and the ligand-based subsets, obtained with different methods and descriptors have been compared. Subsets selected using 2D-fingerprints have been evaluated using B-cut metrics (Figures 5-9, 12 and 13) and the subsets selected by B-cut metrics have been evaluated with 2D-fingerprints (Table 12). The result from the UNITY query has also been compared to the selected subsets using both 2D-fingerprints and B-cut metrics. The resulting subsets from the UNITY query and Protocols 1a and 1b are clustered in the same area using the B-cut metrics Gasteiger charges and polarisability. This indicates that the distributions of the partial charges and the hydrophobicity on the compounds are relevant descriptors for the target protein. Apparently, there are two “islands of activity” in this descriptor space. Using the scale in Figures 5-9, 12 and 13 one of the activity islands (Island 1) is centred at about co-ordinates (4,2) and the other one (Island 2) at about (5,6). Island 1 is occupied by the carbohydrate ligands (the NAG molecules included) and also by some molecules with aromatic rings. Mainly molecules with aromatic rings occupy Island 2 and there are no carbohydrates in this island. Protocols 1a and 1b (the *diversity-based* methods) have started from all over the space and converged to Island 2 whereas Protocols 2a and 2b (the *ligand-based* methods) have both started and ended in Island 1. The ranking energies from FlexX suggest that Island 2 corresponds to higher affinity (a better local energy minimum) than Island 1. An important result from this study is then the discovery of a new activity island for lysozyme, which may contain molecules with higher affinity than the natural substrate. Within this island it

might as well be possible to find inhibitors not only to lysozyme but also to other glycosyl hydrolases with similar active site. This result also suggest that one could reduce the mother database to cover only the two small areas in these descriptors and perhaps even dock all compounds within these areas.

The mean Tanimoto coefficient to the closest neighbour (a measure of the database diversity using 2D-fingerprints) was calculated for all subsets (Table 4). The subsets selected using diversity methods have low mean Tanimoto coefficients (0.26 and 0.51), the DVS *distance-based* subset is diverse also in the 2D-fingerprint descriptors. The subsets selected using similarity methods have high mean Tanimoto coefficients (0.86 and 0.91) *i.e.* the molecules within the subsets are similar. The mean Tanimoto coefficients are not comparable though when the sizes of the subsets are different.

Protocol 1b, *i.e.* the iterative process starting with the diverse subset selected by DVS' distance-*based* algorithm combined with FlexX docking, seems to be a useful procedure judging from the lower docking energies. The hits are also clustered within the same area in B-cut descriptor space as the molecules from the UNITY query. However, it is possible to miss some compounds when the start set is relatively small, and the time needed would increase considerably if a larger number of molecules had to be docked. A much larger subset than the *distance-based* subset above must be selected if the goal is to cover the entire property space of the mother database. It is commonly assumed that the mean Tanimoto coefficient should be 0.85 for a subset to cover the property space of its mother database, to be compared with 0.51 for the DVS *distance-based* subset (Table 4).

Since the UNITY flexible search is not as complex as the FlexX docking, it is possible to search through a much larger set of compounds with this procedure. The UNITY flexible search is suitable for initial virtual screening while FlexX docking is more useful when the number of compounds have been reduced.

Only one common compound was found with the DVS/FlexX iterative method (Protocol 1) and the UNITY flexible search. Probably some compounds included in the results from the iterative method timed out in the UNITY flexible search, because of the time limit. Since only one common compound was found with these methods it would probably be better to use both methods simultaneously.

3.9 Verifying the results

There are several techniques available to experimentally investigate the binding affinities to the protein target. One method is Nuclear Magnetic Resonance (NMR) spectroscopy. NMR is a sensitive method for detecting binding of small molecules to proteins and the method is being used more frequently as a high-throughput screening method to identify ligands (Moore, 1999). It is also possible to detect the binding site on the protein using NMR. Another method to assess binding affinities is by biosensor-based surface plasmon resonance (SPR), which makes it possible to perform kinetic measurements of the ligand binding (Jönsson *et al*, 1991; Ueda *et al*., 1998). Mass spectrometry (MS) is another alternative and it has the possibility to screen a large number of compounds simultaneously. The receptor is trapped in solution by an ultrafiltration membrane, a pulse of compounds in solution are passed through, the unbound compounds are washed away and the bound ligands are identified using MS by mass to charge ratio (m/z) (Swali *et al*., 1999). Crystals of hen egg-white lysozyme are obtained very easily which makes X-ray crystallography an attractive alternative for actually testing if the compounds bind, by analysing difference Fourier electron density maps.

4. Acknowledgements

I am really grateful for the insight into the area of Computational Chemistry that I have received thanks to the enthusiasm from my supervisor *Dr. Enrique Carredano*. I have always been welcome with questions and my computer problems have been solved almost immediately. *Enrique* has also put in a lot of time in helping me with this report.

I also want to thank all my friends here at Amersham Pharmacia Biotech for making my time here pleasant, especially my room (and corridor) mates *Jelena Vasic* and *Mathias Norrman*.

Many thanks also to my family for being supportive during the whole time of the study at the Molecular Biotechnology programme, *Malin, Tobias* and *Peter,* and also my *mother* and my *father*.

5. References

Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J Mol Biol* **112**:535-42.

Bissantz C, Folkers G, Rognan D. (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem*. **43**: 4759-4767.

Blake CC, Johnson LN, Mair GA, North AC, Phillips DC, Sarma VR. (1967a) Crystallographic studies of the activity of hen egg-white lysozyme. *Proc R Soc Lond B Biol Sci*. **167**:378-388.

Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR. (1965) Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* **206**:757-761.

Blake CC, Mair GA, North AC, Phillips DC, Sarma VR. (1967b) On the conformation of the hen egg-white lysozyme molecule. *Proc R Soc Lond B Biol Sci*. **167**:365-377.

Cheetham JC, Artymiuk PJ, Phillips DC. (1992) Refinement of an enzyme complex with inhibitor bound at partial occupancy. Hen egg-white lysozyme and tri-N-acetylchitotriose at 1.75 A resolution. *J Mol Biol*. **224**:613-628.

Conolly ML. (1983) Analytical Molecular Surface Calculation. *J. Appl. Cryst*. **16**:548-558

Creighton TE. (1997) Proteins (2nd edition). WH Freeman and Company. New York

Dorsey BD, Levin RB, McDaniel SL, Vacca JP, Guare JP, Darke PL, Zugay JA, Emini EA, Schleif WA, Quintero JC, *et al*. (1994) L-735,524: the design of a potent and orally bioavailable HIV protease inhibitor. *J Med Chem* **37**:3443-3451.

Duncia JV, Carini DJ, Chiu AT, Johnson AL, Price WA, Wong PC, Wexler RR, Timmermans PB. (1992) The discovery of DuP 753, a potent, orally active nonpeptide angiotensin II receptor antagonist. *Med. Res. Rev*. **12**:149-191.

Duncia JV, Chiu AT, Carini DJ, Gregory GB, Johnson AL, Price WA, Wells GJ, Wong PC, Calabrese JC, Timmermans PB. (1990) The discovery of potent nonpeptide angiotensin II receptor antagonists: a new class of potent antihypertensives. *J Med Chem* **33**:1312-1329.

Glen RC, Martin GR, Hill AP, Hyde RM, Woollard PM, Salmon JA, Buckingham J, Robertson AD. (1995) Computer-aided design and synthesis of 5-substituted tryptamines and their pharmacology at the 5-HT1D receptor: discovery of compounds with potential anti-migraine properties. *J. Med. Chem*. **38**:3566-3580.

Gruneberg S, Wendt B, Klebe G. (2001) Subnanomolar inhibitors from computer screening: a model study using human carbonic anhydrase II. *Angew Chem Int Ed Engl.* **40**:389-393.

Hadfield AT, Harvey DJ, Archer DB, MacKenzie DA, Jeenes DJ, Radford SE, Lowe G, Dobson CM, Johnson LN. (1994) Crystal structure of the mutant D52S hen egg white lysozyme with an oligosaccharide product. *J Mol Biol*. **243**:856-872.

Hansson T, Marelius J, Aqvist J. (1998) Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **12:**27-35.

Holloway MK and Wai JM. (1995) Structure-based design of human immunodeficiency virus-1 protease inhibitors. Ch.3 Computer-aided molecular design. Reynolds

Howell PL, Warren C, Amatayakul-Chantler S, Petsko GA, Hajdu J. (1992) Activity of crystalline turkey egg white lysozyme. *Proteins* **12**: 91-99.

Johnson LN and Phillips DC. (1965) Structure of some crystalline lysozyme-inhibitor complexes determined by X-ray analysis at 6 Angstrom resolution. *Nature* **206**:761-763.

Jönsson U, Fagerstam L, Ivarsson B, Johnsson B, Karlsson R, Lundh K, Lofas S, Persson B, Roos H, Ronnberg I, *et al*. (1991) Real-time biospecific interaction analysis using surface plasmon resonance and a sensor chip technology. *Biotechniques* **11**:620-627.

Kaldor SW, Kalish VJ, Davies JF 2nd, Shetty BV, Fritz JE, Appelt K, Burgess JA, Campanale KM, Chirgadze NY, Clawson DK, Dressman BA, Hatch SD, Khalil DA, Kosa MB, Lubbehusen PP, Muesing MA, Patick AK, Reich SH, Su KS, Tatlock JH. (1997) Viracept (nelfinavir mesylate, AG1343): a potent, orally bioavailable inhibitor of HIV-1 protease. *J Med Chem* **40**:3979-3985.

Kawakami Y, Inoue A, Kawai T, Wakita M, Sugimoto H, Hopfinger AJ. (1996) The rationale for E2020 as a potent acetylcholinesterase inhibitor. *Bioorg Med Chem* **4**:1429-1446.

Keenan RM, Weinstock J, Finkelstein JA, Franz RG, Gaitanopoulos DE, Girard GR, Hill DT, Morgan TM, Samanen JM, Peishoff CE, *et al*. (1993) Potent nonpeptide angiotensin II receptor antagonists. 2. 1-(Carboxybenzyl)imidazole-5-acrylic acids. *J Med Chem* **36**:1880-1892.

Mastryukov VS, Chen K, Yang LR and Allinger NL. (1993) Molecular mechanics calculations (MM3) on bicycloalkyl hydrocarbons. *Theochem* **99**:199-204.

Moore JM. (1999) NMR screening in drug discovery. *Curr Opin Biotechnol* **10**:54-8

Pearlman RS, Smith KM. (1998) Novel metrics and validation of metrics for chemical diversity. *Alfred Benzon Symp.* **42**:165-185.

Pearlman RS. "Concord User's Manual", distributed by Tripos Inc., 1699 South Hanley Rd., St Louis, Missouri, 63144, USA

Pötter T. and Matter H. (1998) Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases. *J. Med. Chem.* **41**:478-488.

Rarey M, Kramer B, Lengauer T, Klebe GJ. (1996) A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J Mol Biol*. **261**:470-489.

Rarey M, Lemmen C, Kramer B, Lengauer CB, Lilienthal M, Sonnenburg F, Zimmermann M. (1999) “FlexS 1.8 User’s Guide”.

Smith L J, Sutcliffe MJ, Redfield C, Dobson CM. (1993) Structure of hen lysozyme in solution. *J Mol Biol.* **229**:930-944.

Steinrauf LK. (1998) Structures of monoclinic lysozyme iodide at 1.6 ångstroms and of triclinic lysozyme nitrate at 1.1 ångstroms. *Acta Crystallogr* D**54**:767-780.

Strynadka NC and James MN. (1996) Lysozyme: a model enzyme in protein crystallography. *EXS.* **75**:185-222. Review.

Swali V, Langley GJ, Bradley M. (1999) Mass spectrometric analysis in combinatorial chemistry. *Curr Opin Chem Biol*. **3**:337-341. Review.

Terrett NK, Bell AS, Brown D and Ellis P. (1996) Sildenafil (Viagra™), a potent and selective inhibitor of type 5 cgmp phosphodiesteras with utility for the treatment of male erectile dysfunction. *Bioorganic & Medicinal Chemistry Letters* **15**:1819-1824.

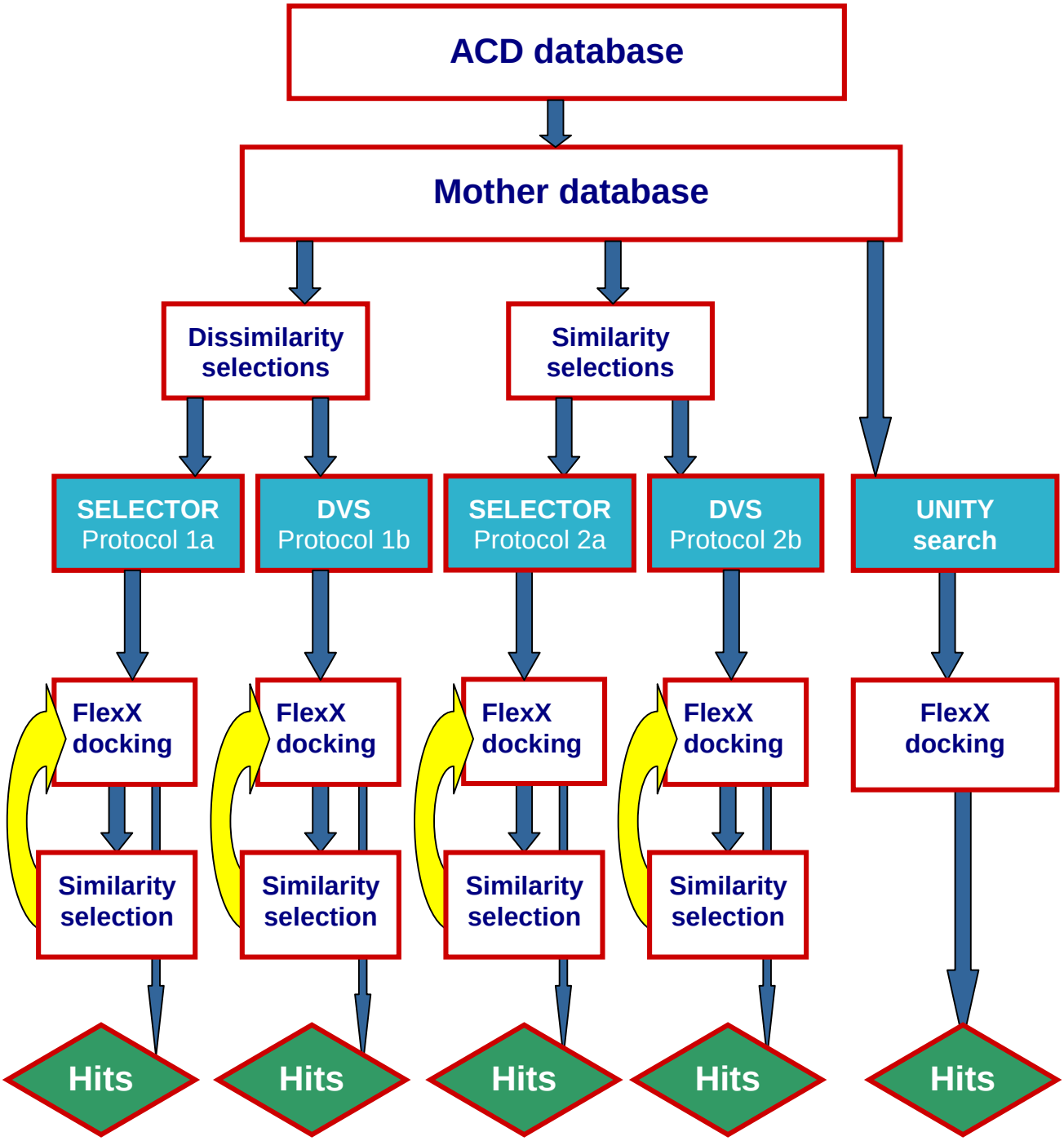
Tripos Inc., 1699 South Hanley Rd., St Louis, Missouri, 63144, USA

Ueda T, Tsurumaru M, Imoto T. (1998) Kinetic measurement of the interaction between a lysozyme and its immobilized substrate analogue by means of surface plasmon resonance. *J Biochem.* **124**:712-716.

Walters WP, Stahl MT, Murcko MA. (1998) Virtual screening – an overview. *DDT.* **3**: 160-178.

Weinstock J, Keenan RM, Samanen J, Hempel J, Finkelstein JA, Franz RG, Gaitanopoulos DE, Girard GR, Gleason JG, Hill DT, *et al*. (1991) 1-(carboxybenzyl)imidazole-5-acrylic acids: potent and selective angiotensin II receptor antagonists. *J Med Chem* **34**:1514-1517.

6. Appendix 1: Flowchart of the methods



7. Appendix 2: PDB structures of hen egg-white lysozyme

Nr	PDB [10]	Resol. (Å)	Ligand	Space group	Rmsd [11] (Å)	First author
1	132l	1.8	+	$P2_12_12_1$	1.4	Rypniewski WR
2	193l	1.33	-	$P4_32_12$	0.5	Vaney MC
3	194l	1.40	-	$P4_32_12$	0.5	Vaney MC
4	1aki	1.50	-	$P2_12_12_1$	1.1	Carter D
5	1at5	1.80	+	$P4_32_12$	0.4	Noguchi S
6	1at6	1.80	+	$P4_32_12$	0.8	Noguchi S
7	1azf	1.80	+	$P4_32_12$	0.9	Lim K
8	1b0d	1.84	+	$P4_32_12$	0.9	Vaney MC
9	1b2k	1.60	+	$P2_1$	1.8	Vaney MC
10	1bvx	1.80	-	$P4_32_12$	0.6	Dong J
11	1bwh	1.80	-	$P4_32_12$	0.5	Dong J
12	1bwi	1.80	-	$P4_32_12$	0.6	Dong J
13	1bwj	1.80	-	$P4_32_12$	0.6	Dong J
14	1dpw	1.64	+	$P4_32_12$	0.8	Weiss MS
15	1dpx	1.65	-	$P4_32_12$	0.9	Weiss MS
16	1f0w	1.90	-	$P2_12_12_1$	1.0	Biswal BK
17	1f10	1.70	-	$P2_12_12_1$	1.2	Biswal BK
18	1hel	1.70	-		0.7	Wilson KP
19	1hew	1.75	+	$P4_32_12$	0.7	Cheetham JC
20	1hsw	2.00	-	$P2_12_12_1$	1.2	Sukumar N
21	1hsx	1.90	-	$P2_12_12_1$	1.0	Sukumar N
22	1kxw	1.96	-	$P4_32_12$	0.8	Motoshima H
23	1kxx	1.71	-	$P4_32_12$	0.7	Motoshima H
24	1kxy	1.79	-	$P4_32_12$	0.7	Motoshima H
25	1lcn	1.63	-	$P2_1$	0.9	Hamiaux C
26	1lkr	1.60	-	$P2_1$	1.0	Steinrauf LK
27	1lks	1.10	-	$P2_1$	1.4	Steinrauf LK
28	1lma	1.75	-	$P2_1$	1.0	Madhusudan R
29	1lsa	1.70	-	$P4_32_12$	1.9	Kurinov I
30	1lsb	1.70	-	$P4_32_12$	1.1	Kurinov I
31	1lsc	1.70	-	$P4_32_12$	0.9	Kurinov I
32	1lsd	1.70	-	$P4_32_12$	1.0	Kurinov I
33	1lse	1.70	-	$P4_32_12$	1.1	Kurinov I
34	1lsf	1.70	-	$P4_32_12$	1.0	Kurinov I

[10] Brookhaven National Laboratory Protein Data Bank code (Bernstein *et al.*, 1977)
[11] Root-mean-square distance from the reference structure 1lzb, based on the main-chain atoms.

Nr	PDB[10]	Resol. (Å)	Ligand	Space group	Rmsd[11] (Å)	First author
35	1lys	1.72	-	$P2_1$	1.3	Harata K
36	1lyz	2.00	-	$P4_32_12$	1.0	Diamond R
37	1lza	1.60	-	$P4_32_12$	0.3	Maenaka K
38	1lzb	1.50	+	$P4_32_12$	0.0	Maenaka K
39	1lzc	1.80	+	$P4_32_12$	0.4	Maenaka K
40	1lzn	1.70	-	P1	1.3	Bon C
41	1lzt	1.97	-	P1	1.3	Hodson JM
42	1rfp	1.75	-	$P4_32_12$	0.7	Motoshima H
43	1uco	2.00	-	$P2_1$	1.0	Nagendra HG
44	1uia	1.76	-	$P4_32_12$	-	Motoshima H
45	1uib	1.76	+	$P4_32_12$	-	Motoshima H
46	1uic	1.95	-	$P4_32_12$	0.7	Motoshima H
47	1uid	1.95	-	$P4_32_12$	0.8	Motoshima H
48	1uie	1.95	-	$P4_32_12$	0.8	Motoshima H
49	1uif	1.85	-	$P4_32_12$	0.7	Motoshima H
50	1uig	1.95	-	$P4_32_12$	0.7	Motoshima H
51	1uih	1.75	+	$P4_32_12$	0.5	Motoshima H
52	1xei	2.10	-	$P2_1$	1.6	Nagendra HG
53	1xej	2.10	-	$P2_1$	1.7	Nagendra HG
54	2lym	2.00	-	$P4_32_12$	0.8	Kundrot CE
55	2lyz	2.00	-	$P4_32_12$	0.9	Diamond R
56	2lzt	1.97	-	P1	1.3	Ramanadham M
57	3lyt	1.90	-	$P2_1$	1.5	Dewan JC
58	3lyz	2.00	-	$P4_32_12$	0.9	Diamond R
59	4lyt	1.90	-	$P2_1$	1.3	Dewan JC
60	4lyz	2.00	-	$P4_32_12$	1.0	Diamond R
61	4lzt	0.95	-	P1	1.3	Walsh MA
62	5lym	1.80	-	$P2_1$	1.0	Rao ST
63	5lyt	1.90	-	$P4_32_12$	1.0	Dewan JC
64	5lyz	2.00	-	$P4_32_12$	1.0	Diamond R
65	6lyt	1.90	-	$P4_32_12$	1.0	Dewan JC
66	6lyz	2.00	-	$P4_32_12$	0.9	Diamond R