



ELSEVIER

Journal of Molecular Structure (Theochem) xx (0000) xxx–xxx

**THEO
CHEM**

www.elsevier.com/locate/theochem

Validation of the SPROUT de novo design program

Jacqueline M.S. Law^{a,b}, David Y.K. Fung^c, Zsolt Zsoldos^{c,*}, Aniko Simon^c,
Zsolt Szabo^c, Imre G. Csizmadia^{a,b,d}, A. Peter Johnson^e

^aDepartment of Chemistry, University of Toronto, Toronto, Ont., Canada M5S 3H6

^bGlobal Institute of Computational Molecular and Materials Science (GIOCOMMS), 1422 Edenrose St., Mississauga, Ont., Canada L5V 1H3

^cSimulated Biomolecular Systems Inc., 135 Queen's Plate Dr, Unit 355, Toronto, Ont., Canada M9W 6V1

^dDepartment of Medical Chemistry, University of Szeged, Dóm tér 8, Szeged 6720, Hungary

^eDepartment of Chemistry, Institute for Computer Applications in Molecular Sciences (ICAMS), University of Leeds, Leeds LS2 9JT, UK

Abstract

The validation of SPROUT was carried out on four receptor–ligand complexes: thrombin–NAPAP, calmodulin (CAM)–AAA, Ras P-21–GDP and dihydrofolate reductase (DHFR)–methotrexate (MTX). These complexes were downloaded from the Brookhaven Protein Data Bank (PDB). For the thrombin–NAPAP complex, two structures very similar to NAPAP were generated. These two structures were similar in 3D structure to NAPAP but contained an extra hexane ring. For CAM–AAA and Ras P-21–GDP, the ligands generated were essentially identical to their original ligands. For DHFR, two ligands, one most similar in 2D structure and one most similar in 3D conformation were found. The successful regeneration of the ligands for each case proves the ability and applicability of SPROUT for designing strongly binding, successful drug candidates. When the program is executed with less restricted constraints, it generates a large number of novel structures that are structurally diverse, making it an ideal tool for de novo design.

© 2003 Published by Elsevier B.V.

Keywords: De novo design; Drug design; Fragment-based exhaustive search; Structure-based; Validation

1. Introduction

The goal of structure-based drug design is to build novel molecular structures or ligands that bind to the receptor of these proteins. Computational methods have been used in the design of novel ligands. Software for structure-based de novo drug design can be divided into three categories: (1) database search techniques, (2) atom based methods that build structures one atom at a time, and (3) fragment joining

techniques that build structures from a library of common molecular fragments.

Hook/MCSS [1] falls into the category of database search techniques. Hook/MCSS exhaustively searches potential binding pockets and docks functional groups to these sites. The functional groups are then linked with templates from a database.

LEGEND [2] is an atom based structure generation program which builds structures one atom at a time. The atom by atom building proceeds as follows. The first atom is placed in a specified distance about a potential hydrogen bonding receptor atom. The subsequent atoms are placed by randomly choosing an atom on the existing ligand, and then placing

* Corresponding author. Tel.: +1-416-741-4263; fax: +1-416-741-5084.

E-mail address: zsolt@simbiosys.ca (Z. Zsoldos).

97 the new atom at a random point on the circle of all
98 possible dihedral angles with fixed bond length and
99 angle. This new atom is assigned an atom type and
100 hybridization. If the new atom occupies a forbidden
101 position on the grid then it is rejected. This procedure
102 terminates when the structure reaches the user defined
103 size.

104 The Allegro [3] program falls into the third
105 category of de novo drug design software. Allegro
106 starts with a pre-computed grid map that classifies the
107 binding zones within the active site. A root atom is
108 chosen and new atoms and fragments grown from this.
109 Growth points are determined and one is randomly
110 selected to form a connection with another randomly
111 chosen atom or functional group. The new atom or
112 group is placed and a complementary score evaluated
113 from the binding zone classification. A Monte Carlo
114 (MC) sampling criteria is used to decide whether this
115 atom or group is to be accepted. If accepted, new
116 growth points are determined and structure generation
117 is repeated from any one of the potential growth
118 points of the molecule. If the distance and bond angle
119 of the growing structure are appropriate, the MC
120 procedure can also spontaneously perform ring
121 closure. This is also called an undirected random
122 search.

123 SPROUT [4] also constructs structures using a
124 fragment joining technique. SPROUT carries out five
125 main functions: (i) locate binding pockets in a
126 receptor, (ii) identify potential interaction sites, (iii)
127 dock molecular fragments to target sites, (iv) generate
128 novel chemical structures by incremental construction
129 from templates and (v) score, sort and cluster the
130 solutions for an efficient means of evaluating the
131 results. In this paper, we present results of validation
132 experiments by using SPROUT to regenerate known,
133 strongly binding inhibitors co-crystallized in their
134 target protein. The idea of these examples is to suggest
135 that if SPROUT can regenerate known inhibitors to a
136 number of targets, this proves the program's ability to
137 generate strongly binding inhibitors. Therefore, any
138 other 'novel' structure generated will also likely be
139 strongly binding. The validation examples presented
140 in Section 3 includes inhibitors of thrombin, calmo-
141 dulin (CAM), Ras P-21 and dihydrofolate reductase
142 (DHFR), all of which are important for mediating and
143 regulating cell cycles.
144

2. Method

An overview of each of the modules will be given
below. However, for a more detailed description of
each module, we refer you to Ref. [4].

2.1. Locating binding pockets

This module offers an automatic tool for finding
potential binding sites within protein structures. In
this way, the program is able to define the receptor site
and the cavity (binding pocket) of a target molecule,
e.g. an enzyme. This module also contains interactive
graphical tools which enable the user to separate co-
crystallized protein ligand complexes and then
generate a cavity and receptor file, respectively,
based on the position of the existing ligand.

2.2. Identification of potential interaction sites

This module identifies potential interaction sites
within a cavity that can be used in de novo design.
These interaction sites define the positions for
potential ligand atoms during structure generation.
This module can detect potential hydrogen bonding,
covalent bonding, metal ions and hydrophobic
interaction sites. An important novel feature of the
method is that it deals with multicentered and
bifurcated hydrogen bonding possibilities.

2.3. Docking of molecular fragments to target site

Before structure generation can begin a set of
target sites, steric boundary, and a set of start
templates are required. This module is designed to
select and orient start templates representing func-
tional groups at the target sites. The functional group
selection is based on the information provided by the
target site identification and takes into account the
distances, relative positions and directions of those
target sites, which are close to each other.

2.4. Generation of structures

This module generates structures that satisfy the
input constraints of the binding pocket. The input is a
set of steric constraints that describe the 3D shape of
the receptor in which the structure must fit. This is

done by joining spacer fragments to the template fragments docked at the target sites and then connecting the resulting partial structures together.

The structure generation phase in SPROUT involves two steps: (1) generating the 3D molecular graphs or template structures that are consistent with the steric constraints of a receptor site; and (2) converting a template structure into a molecular structure with different atom and bond types consistent with the desired electrostatic complementarity.

2.5. Clustering and sorting of results

This module enables the user to cluster/sort the structures generated in SPROUT, into sets using a list of parameters specified by the user. In this way, unwanted structures can be identified and discarded leaving only structures that meet the requirements of the user.

The solutions are ranked using a scoring function that assesses hydrogen bonds, van der Waals interactions, rotatable bonds and hydrophobic surface contact area. The sorting of results can be done by various criteria such as number of gauche interactions, stereo centers and ring fusions. An additional function of this module is hetero-atom substitution in which generic atoms are replaced by specific atoms to fulfill interaction type criteria (e.g. hydrogen bond donor) at target sites, leading to final solutions which may have better binding score or stability score.

3. Results and discussion

3.1. Thrombin

Thrombin is a serine protease that mediates the blood clotting cascade and converts fibrinogen into fibrin through hydrolysis. The crystal structure of thrombin complexed with NAPAP from the Protein Data Bank (PDB code 1ets) was used for the first validation of SPROUT. Thrombin has three binding sites, one of which has an aspartic acid (Asp) residue that interacts with a positively charged part of a ligand. The other two sites are hydrophobic sites, one of which allows the binding of an aromatic moiety of a ligand. Studies have shown that the ligand NAPAP

exhibits quite a strong binding affinity towards this pocket [5].

The hydrogen bond donor site OD1 ASP H189, hydrogen bond acceptor site N GLY H216 and manually added hydrophobic spheric sites were chosen on the receptor for the interactions sites in the generation of a ligand. These sites are shown in Fig. 1. The spheric sites were very important because they helped in the regeneration of a structure similar to NAPAP. In addition, they also model the hydrophobic interactions between protein and ligand. Without the spheric target sites, SPROUT would generate a large diverse set of structures, many of them very different from NAPAP.

Starting fragments similar in structure to NAPAP were docked into the target sites with different positions and conformations as shown in Fig. 1. Note that for the fragment with the NHSO_2 group, two sites were used to dock the fragment. These two sites anchored the fragment such that the methyl end and the NHSO_2 would be oriented properly.

Two structures found to closely resemble NAPAP are shown in Fig. 2. However, they both have an extra six-member ring, hexane, which may result in this new structure being more bulky. However, it seems to

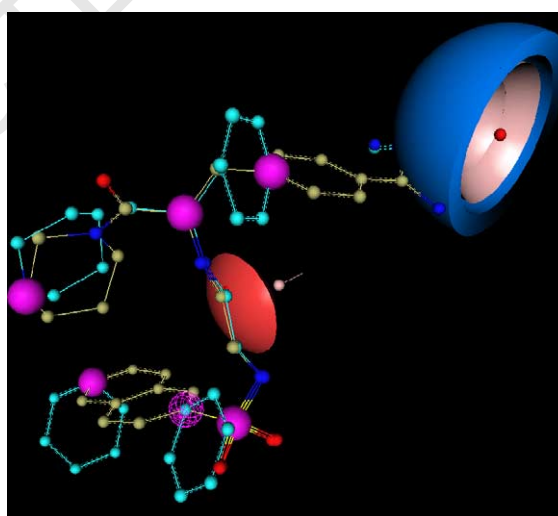


Fig. 1. The target sites chosen for regeneration of NAPAP. The generic spheric sites are in pink and the hydrogen bond donor and acceptor sites are the blue and red semi-spheres, respectively. A comparison of the fragments docked at the target sites (light blue) and NAPAP (brown) are also shown.

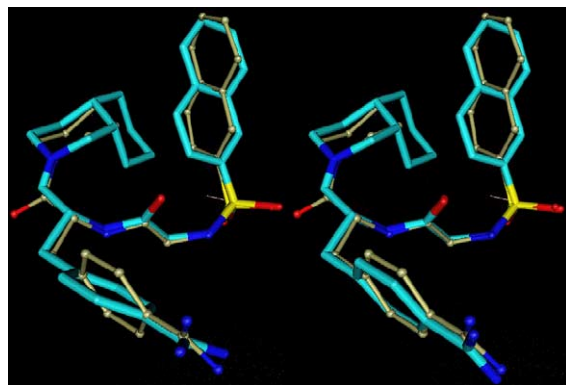


Fig. 2. Two SPROUT generated structures (blue) resembling the original NAPAP ligand (brown).

form additional hydrophobic interactions between the ligand and the receptor.

3.2. Calmodulin

CaM is known to control processes such as cell division and gene expression [6]. To activate different proteins, CaM complexes with Ca^{2+} , then binds onto the polypeptides of different receptor proteins and induces conformational change [7].

Many studies have been carry out on ligands such as *N*-(3,3-diphenylpropyl)-*N'*-[1-*R*-(3,4-bis-butoxyphenyl)ethyl]-propylendiamine (AAA) that mimic the receptor peptides that bind to CaM. The CaM–AAA complex (PDB code 1qiw) was used because AAA has a high affinity for CaM in the presence of Ca^{2+} atoms. Without Ca^{2+} , AAA would not bind to CaM at all.

In total, nine sites were selected. One site is a hydrogen bond donor site OE1 GLU-B11 and eight of them were hydrophobic sites represented by the spheric sites. The sites that were chosen are shown in Fig. 3. Since the biphenyl groups can interact with the hydrophobic cleft, the sites must be put near those clefts to ensure that the phenyl groups can be grown. Fig. 3 also shows the fragments docked onto the target sites. The blue structures are the fragments docked and the brown structure is AAA.

In the structure generation module, these fragments were joined using template fragments similar to those found in AAA. There were many solutions found, however, one solution had a very close resemblance to the original ligand shown in Fig. 4.

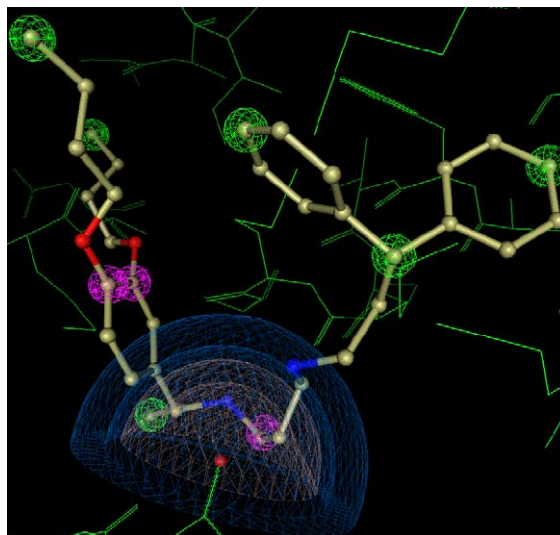


Fig. 3. Fragments docked to target sites of calmodulin. The hydrophobic sites are the green spheres, generic spheric sites in pink and the hydrogen bond donor sites is the blue semi-sphere.

3.3. Ras P-21 and GDP complex

P21 is an oncogene found in the Ras family known to be related to different human tumors [8]. Under normal circumstances, this oncogene is inactivated when GTP is hydrolyzed to GDP by a GAP protein [9]. However, a point mutation in a base would result in GAP being unable to bind to and hydrolyze the P21–GTP complex to P21–GDP [10]. In turn, the Ras protein would always be activated and would become an oncogene [11].

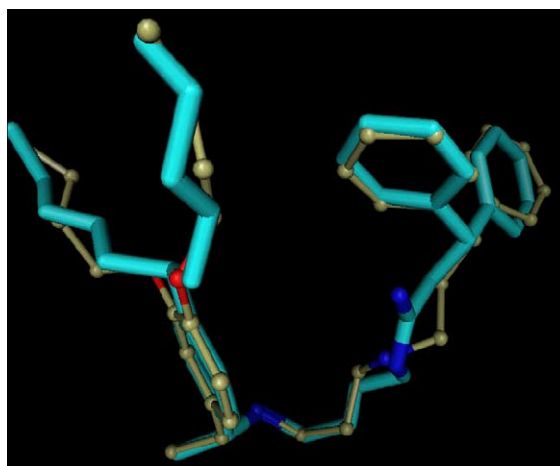


Fig. 4. The final structure (light blue) that resembles AAA (brown).

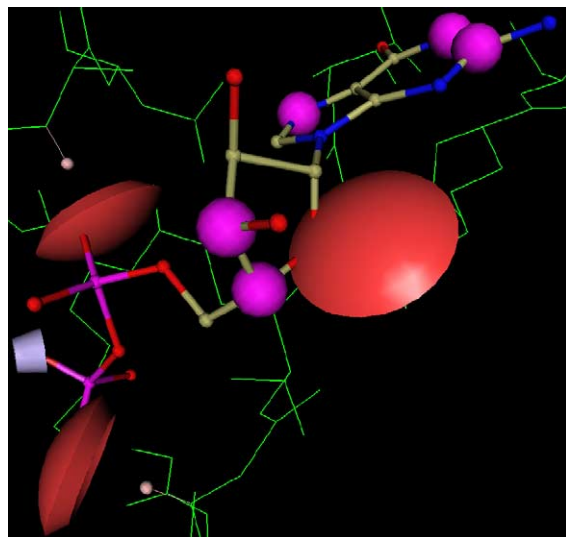


Fig. 5. Sites in Ras P-21 chosen for the docking of fragments.

The Ras P-21 protein complexed with GDP (PDB code 1q21) was studied. The target sites satisfying the ligand include a hydrogen bond acceptor site on the pentose moiety of guanidine (N Gly-13), two hydrogen bond acceptor sites (NZ Lys-117 and N ALA-18), each located at one of the phosphate groups of the ligand, and a divalent metal ion site (Mg Mg-173) at the terminal phosphate group shown as a silver cylinder in Fig. 5. The metal ion is an important binding site for the GDP molecule. Five spheric sites were also chosen to guide the extension and growth of

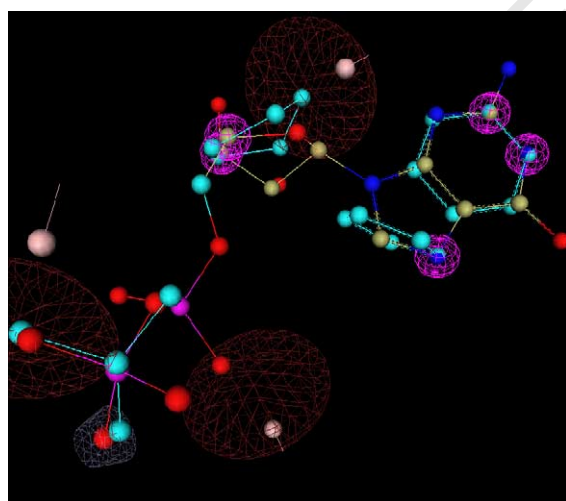


Fig. 6. Fragments chosen for docking at the target sites of Ras P-21.

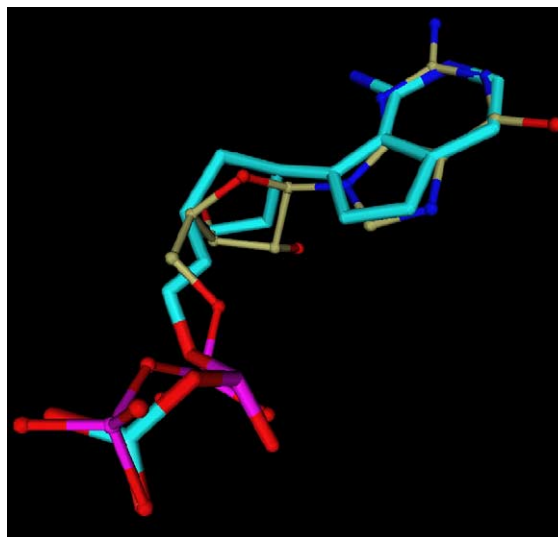


Fig. 7. The solution (light blue) mimicking the GDP ligand (brown).

this molecule. Fig. 5 shows the target sites selected for this molecule. The fragments chosen as starting points are shown in Fig. 6. The three sites used for the pentose sugar ensured that the five-member ring would have the correct geometry. This method also ensures that the geometry of this molecule would also stay unchanged when the joining fragments are added.

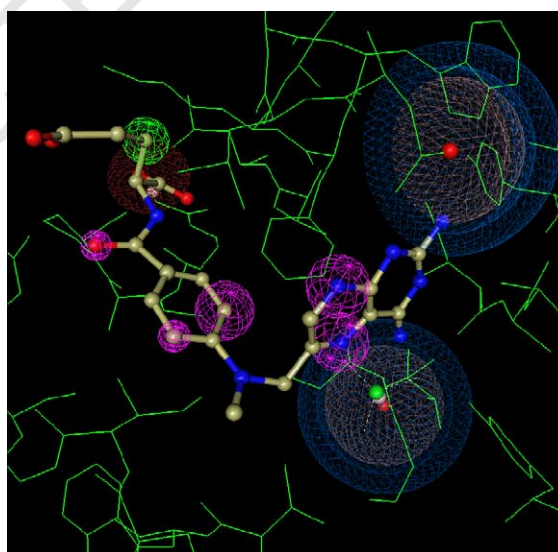


Fig. 8. The target sites chosen in DHFR. The hydrophobic site is the green sphere, generic spheric sites in pink and the hydrogen bond donor and acceptor sites are the blue and red semi-spheres, respectively. The original ligand MTX is in brown.

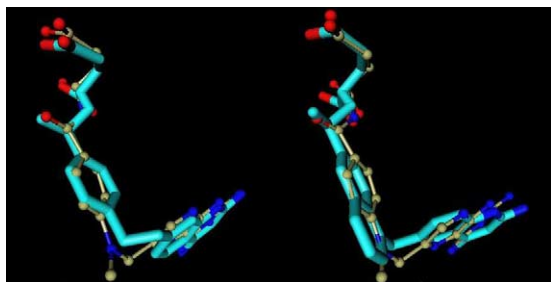


Fig. 9. Structures most similar to MTX (brown). The ligand on the left is most similar in 2D structure and the ligand on the right is most similar in 3D structure.

Fig. 7 shows the final ligand structures mimicking GDP.

3.4. Dihydrofolate reductase and methotrexate

DHFR is essential for producing tetrahydrofolate by the NADPH facilitated reduction of dihydrofolate [12]. Tetrahydrofolate is essential in the synthesis of metabolites in cells. In cancerous cells, the high levels of DHFR are essential for the proliferation of cells. The methotrexate (MTX) ligand can inhibit DHFR and can thus serve as both an anticancer and antitumour drug.

The crystal structure of DHFR complexed with MTX (PDB code 4dfr) was used as the next model for validation of SPROUT. In the DHFR receptor, a number of hydrogen bonding sites were chosen. They included the hydrogen bond donor sites OD1 ASP-A27 and O Ile-A94, the hydrogen bond acceptor site NH2 Arg-A57, as well as a number of hydrophobic sites on the two aromatic rings and at the end of the chain. The hydrogen bonding sites were used such that the newly designed ligand would have a relatively good hydrogen bonding interaction with the active sites of the receptor cleft. Fig. 8 shows the target sites chosen.

In this model, two structures similar to MTX were found. Although only two results are shown, other hetero-atom substitutions on atoms at other positions are possible. An image of the two most similar molecules are shown in Fig. 9.

4. Conclusions

SPROUT provides an exhaustive, deterministic solution for de novo ligand design. This approach

can guarantee to find the best solution with the defined constraints. This is a very valuable and unique feature, differentiating SPROUT from the numerous de novo tools that apply random techniques or artificial limitations imposed by database searches.

The SPROUT de novo design software was used to regenerate known, strongly binding ligands for thrombin, CAM, Ras P-21 and DHFR. The ligands structurally and conformationally resembling the co-crystallized X-ray ligand were successfully regenerated for all these receptors. The success of SPROUT in this validation demonstrates its ability to regenerate known strongly binding inhibitors suggesting that it has great potential in deriving unknown, novel ligands for other target receptors.

Acknowledgements

This study was made possible by the Scientific Research and Experimental Development grant from the National Research Council of Canada and the Global Institute of Computational Molecular and Materials Science (GIOCOMMS), Toronto, Ont., Canada. One of the authors (IGC) wishes to thank the Ministry of Education for a Szent-Györgyi Visiting Professorship.

References

- [1] (a) A. Miranker, M. Karplus, *Protein Struct. Func. Gen.* 11 (1991) 29.
(b) A. Caflish, A. Miranker, M. Karplus, *J. Med. Chem.* 36 (1993) 2142.
(c) M.B. Eisen, D.C. Wiley, M. Karplus, R.E. Hubbard, *Protein Struct. Func. Gen.* 19 (1994) 199.
- [2] (a) Y. Nishibata, A. Itai, *Tetrahedron* 47 (1991) 8985.
(b) Y. Nishibata, A. Itai, *J. Med. Chem.* 36 (1993) 2921.
- [3] R.S. Bohacek, C. McMartin, *J. Am. Chem. Soc.* 116 (1994) 5560.
- [4] (a) V.J. Gillet, A.P. Johnson, P. Mata, S. Sike, P. Williams, *J. Comput.-Aided Mol. Des.* 7 (1993) 127.
(b) V.J. Gillet, W. Newell, P. Mata, G.J. Myatt, S. Sike, Z. Zsoldos, A.P. Johnson, *J. Chem. Inf. Comp. Sci.* 34 (1994) 207.
(c) P. Mata, J. V., A.P. Gillet, J. Johnson, G.J. Lampreia, S. Myatt, A.L. Sike, Stebbings, *J. Chem. Inf. Comp. Sci.* 35 (1995) 479.

577	[5] D.W. Banner, P. Hadvar, <i>J. Biol. Chem.</i> 266 (1991) 20085.	[9] A. Wolfman, I.G. Macara, <i>Science</i> 248 (1990) 67.	625
578	[6] A. Crivici, M. Ikura, <i>Ann. Rev. Biophys. Biomol. Struct.</i> 24 (1995) 85.	[10] J. Downward, R. Riehl, L. Wu, R.A. Weinberg, <i>Proc. Natl Acad. Sci. USA</i> 87 (1990) 5998.	626
579	[7] N. Takuwa, W. Zhou, Y. Takuwa, <i>Cell. Signal.</i> 7 (1995) 93.	[11] T. Schweins, K. Scheffzek, R. Auheuer, A. Wittinghofer, <i>J. Mol. Biol.</i> 266 (1997) 847.	627
580	[8] A.M. deVos, L. Tong, M.V. Milburn, P.M. Matias, J. Jancarik, S. Noguchi, S. Nishimura, K. Muira, D. Ohtsuka, S. Kim, <i>Science</i> 239 (1988) 888.	[12] V.M. Reyes, M.R. Sawaya, K.A. Brown, J. Kraut, <i>Biochemistry</i> 34 (1995) 2710.	628
581			629
582			630
583			631
584			632
585			633
586			634
587			635
588			636
589			637
590			638
591			639
592			640
593			641
594			642
595			643
596			644
597			645
598			646
599			647
600			648
601			649
602			650
603			651
604			652
605			653
606			654
607			655
608			656
609			657
610			658
611			659
612			660
613			661
614			662
615			663
616			664
617			665
618			666
619			667
620			668
621			669
622			670
623			671
624			672