In Silico Drug Discovery: Solving the "Target-rich and Lead-poor" Imbalance Using the Genome-to-drug-lead Paradigm

YP Pang[1]

Advances in genomics, proteomics, and structural genomics have identified a large number of protein targets. Virtual screening has gained popularity in identifying drug leads by computationally screening large numbers of chemicals against experimentally determined protein targets. In that context, there continues to be a "target-rich and lead-poor" imbalance, reflecting an insufficiency of chemists pursuing drug discovery in academia, the challenge of engaging more chemists in this area of research, and a paucity of available protein target structures. This imbalance in manpower and structural information can be ameliorated, in part, by adapting a "genome-to-drug-lead" approach, in which chemicals can be virtually screened against computer-predicted protein targets, within the context of the US National Science Foundation's petascale computing initiative. This approach offers a solution to reduce manpower requirements for more chemists to experimentally search for drug leads, which represent one of the greatest limitations to drug discovery and better exploits the extensive availability of drug targets at the gene level, ultimately improving the success of moving discoveries from the laboratory to the patient.

TARGET-RICH AND LEAD-POOR

The completion of the Human Genome Project in 2003 and recent advances in proteomics and the Structural Genomics Initiative have identified a large number of human proteins as drug targets whose activities can be specifically affected by traditional small organic molecules (chemicals).[1–3] The human genome has advanced our understanding of the scientific basis of individual variations, and those variations caused by single-nucleotide polymorphism have further increased the number of potential drug targets to an estimated 5,000 (http://www.bio-itworld.com/archive/100902/firstbase.html). At the same time, the number of chemicals generated by traditional and contemporary approaches has increased dramatically. In theory, there could be as many as 10^{47} quadrillion chemicals that can be made to interact with human protein targets.[4]

To test this myriad of chemicals, computational screening (virtual screening) can be pursued by iteratively docking each chemical into the active site of a protein target to identify drug leads.[5–8] Identification is based on the evaluation of the fitness between the two molecules in terms of their shapes and charges. Virtual screening has gained popularity in identifying drug leads with potencies of less than 100 μM by screening chemicals against a protein structure determined by single-molecule X-ray crystallography or nuclear magnetic resonance spectroscopy. In theory, virtual screening is *scalable* computationally and could yield the desired balance between available therapeutic targets and the identification of drug lead compounds. However, there remains a "target-rich and lead-poor" imbalance. Why?

One obvious reason is that, relative to biologists, there is a paucity of organic/medicinal chemists in academia who are supported to experimentally identify drug leads. This situation could be ameliorated by policies directing support to this endeavor. However, a change in policy to support more chemists pursuing drug research will not immediately correct the imbalance. Skilled organic/medicinal chemists are expensive and require time for adequate training, reflecting their need for *tacit*, rather than *explicit*, knowledge to *create* chemicals as potential drugs. It typically takes 4–6 years of training to acquire the tacit knowledge required for drug design and organic synthesis. Therefore, in the context of the shortage of organic/medicinal chemists to search experimentally for drug candidates, computational approaches offer a

[1]Computer-Aided Molecular Design Laboratory, Mayo Clinic College of Medicine, Rochester, Minnesota, USA. Correspondence: YP Pang (pang@mayo.edu)
doi:10.1038/sj.clpt.6100030

solution to better balance the availability of therapeutic targets and the identification of drug leads.

Technically, virtual screening appears to offer a viable solution to optimize therapeutic drug candidate discovery. A one-teraflops computer is able to perform 31.536×10^{18} float point operations per year. In a highly simplified scenario, a dedicated one-teraflops computer in a year could screen 200 million chemicals for each of the 5,000 protein targets, in a year. This scenario assumes that it takes 31.536×10^{6} float point operations to screen one chemical against a protein target at a resolution of 1.0-Å translational increment in a $3 \times 3 \times 3$-Å^3 docking box and 10° of arc rotational increment in the *x*, *y*, and *z* directions.[9] According to the US National Science Foundation's petascale science and engineering initiative (http://www.nsf.gov/pubs/2005/nsf05625/nsf05625.htm), in 2010 a one-petaflop computer will perform 31.536×10^{21} float point operations, with a theoretical capacity to screen 200 *billion* chemicals for each of the 5,000 protein targets, in a year. Although 200 billion chemicals are a small fraction of 10^{47} quadrillion chemicals, this is more than the number of chemicals tested for the development of any clinical therapeutic developed to date.

Although this computational approach clearly offers a solution to better balance the availability of drug targets and the identification of therapeutic lead candidates in the current drug discovery/development paradigm, there are limitations to this virtual screening model. Many human protein targets, especially those with variations caused by single-nucleotide polymorphism, currently do not have three-dimensional (3D) structures defined. This lack of 3D protein structure for targets prohibits the application of virtual screening to identify lead drug candidates. In that context, a new approach is required.

THE GENOME-TO-DRUG-LEAD APPROACH

Whereas several approaches can be employed to define 3D protein structures, the primary method is to experimentally determine structures of globular proteins bearing unique folds through the Structural Genomics Initiative.[3] A complementary method is to predict 3D protein structures from their sequences. By combining improved low- and high-resolution conformational sampling methods, 1.5-Å-resolution structure prediction has been achieved for small protein domains with less than 85 amino acids.[10]

This advance and our own protein modeling experience described below suggest that virtual screening can be expanded to screen chemicals against protein targets whose active site-containing domain or subdomain is predicted computationally, an ambitious approach we term "genome-to-drug-lead".[11] To illustrate feasibility, we built a dedicated 1.1 teraflops computer (**Figure 1**) to run multiple molecular dynamics simulations (MMDSs) in parallel. The stochastic sampling of protein conformations achieved by MMDSs is more efficient than sampling by a single long molecular dynamics simulation.[12–17] The efficiency of the stochastic sampling is demonstrated by MMDSs of the ubiquitin E2 variant domain of human tumor susceptibility gene 101 protein in complex with a peptide ligand in explicit water.[18] Here, the MMDSs comprise 200 different 10-ns molecular dynamics simulations (2×10^{6} snapshots) of the complex for which nuclear magnetic resonance data are available.[18] The trajectories obtained during the first 5 ns period of the MMDSs reproduce ~92% of the protein–protein nuclear Overhauser effects and ~85% of the protein–peptide nuclear Overhauser effects (YP Pang and P Dasgupta, unpublished data). Given this sampling efficiency, MMDSs could refine a low-resolution protein domain, which is readily obtained from homology modeling or threading, to a high-resolution protein domain. For example, MMDSs refined a homology model, provided by the Protein Structure Prediction Centre (http://predictioncenter.org/caspR/), to a computer model that was nearly identical to the corresponding crystal structure (Protein Data Bank ID: 1XE1). Relative to the 1XE1 crystal structure, the alpha carbon root mean square deviation of the computer model was 1.7 Å, whereas the alpha carbon root mean square deviation of the homology model was 4.6 Å (**Figure 2**, unpublished work of Pang).

Figure 1 Kibbutz100, a homemade 1.1 teraflops supercomputer for virtual screening and multiple molecular dynamics simulations.

In the context of this advanced performance, we applied homology modeling and MMDSs to predict a 3D model of a chymotrypsin-like cysteine proteinase (CCP) from a severe acute respiratory syndrome-associated coronavirus.[11,17] CCP is an ideal drug target for treating severe acute respiratory syndrome viral infection because it is required for viral replication and transcription. Here, 200 different molecular dynamics simulations of monomeric CCP in explicit water (4.0 ns for each simulation with a 1.0 fs time step and different initial velocities) were executed to refine the homology model.[11,17] Then, we screened 361,413 chemicals against the CCP model refined by MMDSs and identified 12 chemicals for antiviral testing. Of the 12 chemicals tested in cell-based inhibition assays, one inhibited the human severe acute respiratory syndrome–coronavirus Toronto-2 strain with a concentration of ligand that produces half of the

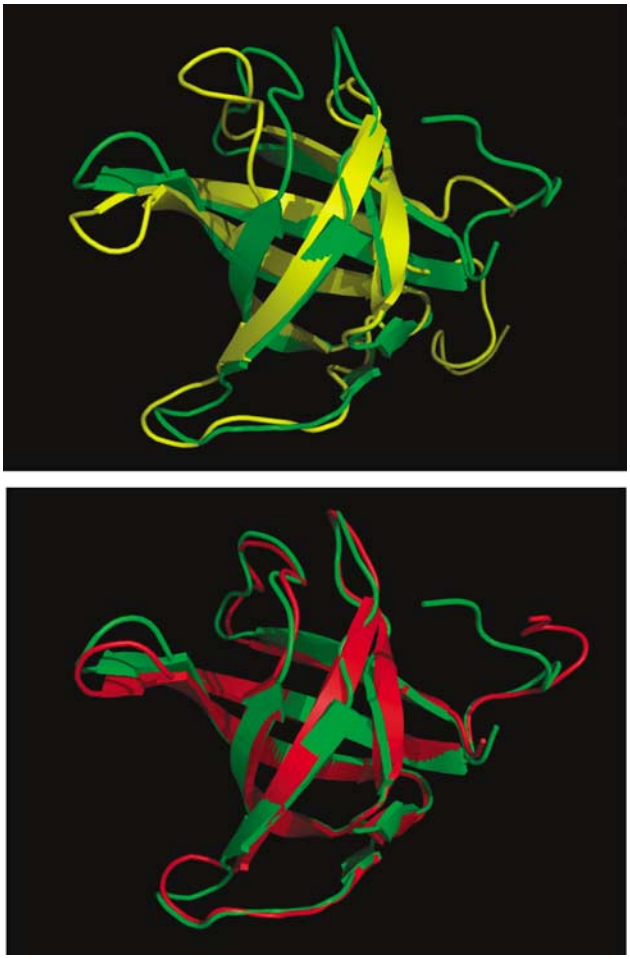

Figure 2 Overlays of a crystal structure (green, Protein Data Bank ID: 1XE1) with a homology model (yellow, CASP6 model T0196TS451_1) and with a model refined from the homology model (red), respectively.

maximum response of 23 μM and four others exhibited 13–17% inhibition at a drug concentration of 32 μM.[11] The most potent inhibitor lead overlays well with a reported substrate fragment ($ATVRLQ^{P1}A^{P1'}$) bound in the active site of CCP (**Figure 3**)[17]. These results demonstrate that, given target information at the gene level only, virtual screening can identify chemicals that penetrate and rescue cells from viral infection. It is noteworthy that this genome-to-drug-lead approach leapfrogs the requirements for experimental determination of protein target structure and cell-free assays to confirm molecular interactions.

Interestingly, CCP exists in a homodimer in which only one of the two monomers is active.[19] Thus, simulation of monomeric CCP may lead to a structure that is not representative of the active CCP. In fact, many protein targets are functional only in a multimeric form. With target information at the gene level only, it is difficult to deduce the precise multimeric form required for the function of the protein target, let alone the challenge of simulating proteins in their multimeric forms. This problem appears to mitigate against the use of the genome-to-drug-lead approach. However, information concerning ternary structure is not

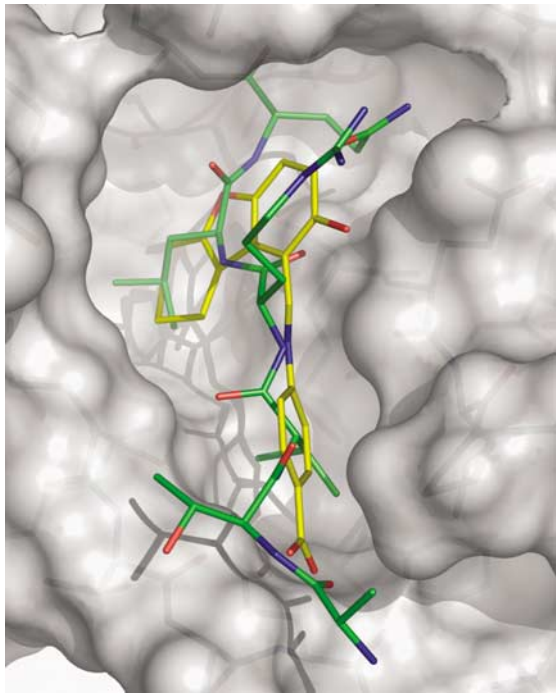

Figure 3 Overlay of the 23 μM inhibitor with a substrate fragment ($ATVRLQ^{P1}A^{P1'}$) bound in the active site of CCP.

required if virtual screening is searching for inhibitors (not activators) of protein targets. Indeed, an inhibitor lead identified from the inactive, monomeric CCP binds to the monomeric CCP, and possibly to the dimeric CCP as well. While binding to dimeric CCP can certainly inhibit CCP, binding to monomeric CCP also can inhibit CCP because the dimer is in equilibrium with the monomer and binding to the monomer can convert the active dimeric CCP to the inactive monomeric CCP. This explains why a 23 μM inhibitor lead was successfully identified using the inactive monomeric CCP in virtual screening.

While virtual screening identified compounds that inhibited CCP activity in cell-based assays, this approach did not empirically validate model predictions by examining direct molecular interactions in cell-free systems. However, a screen using the same CCP model but contracted by 12% (**Figure 4**) failed to identify the 23 μM inhibitor (**Figure 5**).[11] Moreover, it identified two weak inhibitors that are structurally very similar to the 23 μM inhibitor.[11] These observations demonstrate that the identification of a drug lead is sensitive to a change of the structure used in virtual screening, implying the interaction of the lead with CCP, and thereby confirms the validity of virtual screening using the inactive monomeric CCP. Further, it demonstrates that leapfrogging the cell-free assay has the advantage of avoiding identification of both toxic inhibitors and inhibitors that have high affinities for CCP but cannot penetrate cells.

The best CCP inhibitor lead has a concentration of ligand that produces half of the maximum response of only 23 μM, which raises the question of its utility. Using the traditional trial-and-error approach, it is certainly difficult to improve

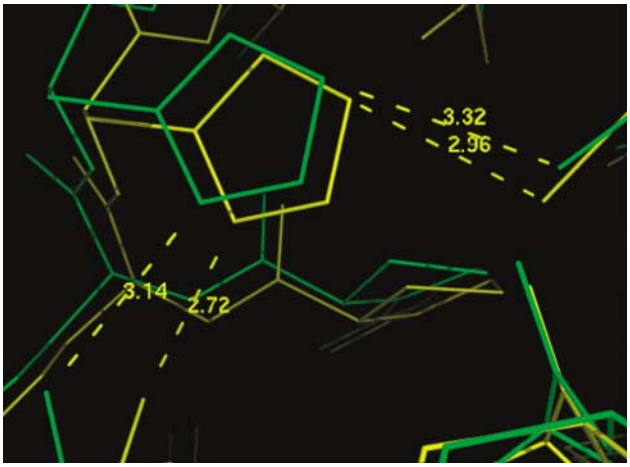

Figure 4 Overlay of CCP's catalytic triad in the original model and the slightly contracted model that was generated by averaging without RMS fit.

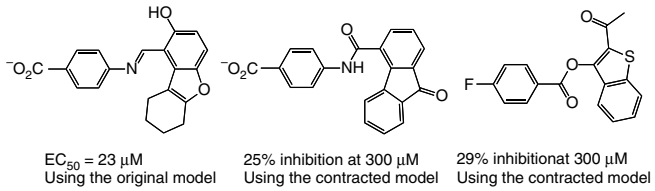

Figure 5 Chemicals identified by virtual screening using the original model and the slightly contracted model.

Figure 6 Enabling breakthrough science with IBM Blue Gene®/L.

the potency of a 100 μM lead by several orders of magnitude. For this reason, the definition of a drug lead is commonly defined as a chemical possessing an inhibitory potency less than 50 μM. However, using MMDSs to guide structural modification, we improved an inhibitor lead of a zinc endopeptidase in botulinum neurotoxin serotype A from 15% inhibition at a drug concentration of 100 μM to 19% inhibition at 2.5 μM.[20] This demonstrates that the 23 μM inhibitor is useful as a drug lead, especially because its potency was determined by a cell-based assay. It also suggests that a drug lead can be re-defined as a chemical possessing an inhibitory potency less than 100 μM.

To further appreciate the value of drug leads obtained from virtual screening, it is worth discussing the goal of virtual screening because this goal may differ among research groups and change over time. In 2000, our goal of virtual screening was to identify a subset of chemicals enriched in active inhibitors[5] based on the gigaflops (10^9 floating point operations per second) computing technology available then. In 2006, we have the same goal, even though 3.8 teraflops (3.8×10^{12} floating point operations per second) computing technology has become available. The goal has not changed because terascale (10^{12} floating point operations per second) computers remain insufficiently fast to identify drug candidates. Drug discovery relies on organic/medicinal chemists who have the tacit knowledge to create chemicals as drug candidates with the aid of fast computers. We do not anticipate that virtual screening can identify *drug candidates* that are ready for preclinical studies. Rather, we expect that virtual screening can offer *drug leads* as building blocks for drug candidates. Virtual screening cannot create new chemicals, but organic chemists can use leads as building blocks to create new chemicals. That is the value of drug leads obtained from virtual screening.

An unusually large computing resource was used to search for CCP inhibitors. Would the genome-to-drug-lead approach be practical for a typical academic research laboratory? Indeed, it is practical for several reasons. First, the cost of a 2.2 Ghz Intel Xeon processor was $800 in 2002, but is $79 in 2006; similarly a 1.0 teraflop computer was $400,000 in 2002, but can be built for less than $50,000 in 2006. Second, to screen 23,426 chemicals (at a resolution of 1.0 Å translation and 10° of arc rotation) for CCP inhibitors, our in-house docking program EUDOC/BLEUDOK[9] can reduce wall-clock time from 242 min by using 396 Xeon processors (2.2 GHz) on a Beowulf cluster to 13 and 7 min by using 2,048 and 4,096 PowerPC-440 processors (700 MHz) on IBM Blue Gene®/L supercomputers (**Figure 6**), respectively (unpublished work of YP Pang, CJ Archer, JS Mcallister, TJ Mullins, RG Musselman, AE Peters, KW Pinnow, BE Smith, BA Swartz, and BP Wallenfelt, unpublished data). Third, according to the US National Science Foundation's petascale science and engineering initiative, in 2010 petascale computers will be made available and the computing technology used for the CCP inhibitor work will become trivial.

A relatively small database of 361,413 chemicals was used to identify CCP inhibitors. However, this success of identifying CCP inhibitor leads from a relatively small library of chemicals does not address the feasibility of docking targets with 200 billion chemicals. In that context, data storage is a common problem in bioinformatics. Do we have enough disk space to store 200 billion chemicals to be screened as discussed in the previous section? The answer is yes. The average disk space to hold one chemical with a molecular weight in a range of 380–420 is 973 bytes, according to the database designed by this author. A well-designed database of 200 billion chemicals is estimated to

take 200 terabytes (2×10^{14} bytes) of disk space. The cost of 200 terabytes of disk space is $100,000 in 2006 and will decrease significantly by 2010.

In summary, given advances in protein structure prediction and the change in computing speed, from gigascale in the past, to terascale in the present, and to petascale in the near future, it is clear that the genome-to-drug-lead approach is feasible and has broad application in drug discovery.

IMPACT ON CLINICAL PHARMACOLOGY AND THERAPEUTICS

The potential impact of the genome-to-drug-lead approach on clinical pharmacology and therapeutics is the increase in the number of drug candidates that can be identified and moved from the laboratory into clinical trials. However, the impact goes further. The genome-to-drug-lead approach permits the docking of one drug candidate against an array of human proteins to predict drug interactions and toxicity, and effectively address individual variations of a drug target caused by single-nucleotide polymorphism for personalized medicine. A slight modification of the genome-to-drug lead approach can dock the computer-identified drug candidate against human serum albumin to predict protein binding and, by extension the distribution, of drug candidates.

CONCLUSION

A large number of drug targets and a paucity of organic/medicinal chemists in academia pursuing drug discovery research have created a "target-rich and lead-poor" imbalance. Skilled organic/medicinal chemists are expensive and take time to train and it is difficult to engage more academic organic/medicinal chemists in drug research to remediate this imbalance in the short term. Virtual screening can successfully identify drug leads against experimentally determined drug target structures. Given current terascale computers, and petascale computers in the near future, virtual screening can be expanded to screen chemicals against target structures predicted from genes by computers, a paradigm termed "genome-to-drug-lead". This approach can reduce the formidable manpower requirements for more chemists to search experimentally for lead compounds, help resolve the imbalance between disease targets and therapeutic agents, and, ultimately, enrich the drug development pipeline to move discoveries from the laboratory into patients.

ACKNOWLEDGMENTS

The author's work described here was supported by the Defense Advanced Research Projects Agency (DAAD19-01-1-0322), the US Army Research Office (DAAD19-03-1-0318), the US Army Medical Research Acquisition Activity (W81XWH-04-2-0001), the National Institutes of Health (5R01AI054574-03 and 5R01GM061300-06), the IBM Blue Gene Life Sciences Center of Excellence, the Mayo Clinic-IBM Center of Excellence, the High Performance Computing Modernization Program of the US Department of Defense, the San Diego Supercomputing Center, the University of Minnesota Supercomputing Institute, the Compaq Medical Sciences Group, the Jay and Rose Phillips Family Foundation, and the Mayo Foundation. The opinions or assertions contained herein belong to the author and are not necessarily the official views of the US Army, the US Department of Defense, or the National Institutes of Health.

CONFLICT OF INTEREST

The author declared no conflict of interest.

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
2. Banks, R.E. *et al.* Proteomics: new perspectives, new biomedical opportunities. *Lancet* **356,** 1749–1756 (2000).
3. Burley, S.K. *et al.* Structural genomics: beyond the Human Genome Project. *Nat. Genet.* **23,** 151–157 (1999).
4. Chait, E.M. Drug discovery – contemporary small molecule drug discovery – Tutorial: stacking the deck in favor of drug-like leads. *Genet. Eng. News* **22,** 34–37 (2002).
5. Perola, E. *et al.* Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *J. Med. Chem.* **43,** 401–408 (2000).
6. Miller, M.A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* **1,** 220–227 (2002).
7. Sousa, S.F., Fernandes, P.A. & Ramos, M.J. Protein–ligand docking: current status and future challenges. *Proteins* **65,** 15–26 (2006).
8. Hattotuwagama, C.K., Davies, M.N. & Flower, D.R. Receptor–ligand binding sites and virtual screening. *Curr Med Chem* **13,** 1283–1304 (2006).
9. Pang, Y.-P., Perola, E., Xu, K. & Prendergast, F.G. EUDOC: A computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comp. Chem.* **22,** 1750–1771 (2001).
10. Bradley, P., Misura, K.M. & Baker, D. Toward high-resolution *de novo* structure prediction for small proteins. *Science* **309,** 1868–1871 (2005).
11. Dooley, A.J., Shindo, N., Taggart, B., Park, J.G. & Pang, Y.-P. From genome to drug lead: identification of a small-molecule inhibitor of the SARS virus. *Bioorg. Med. Chem. Lett.* **16,** 830–833 (2006).
12. Caves, L.S.D., Evanseck, J.D. & Karplus, M. Locally accessible conformations of proteins – multiple molecular dynamics simulations of crambin. *Protein Sci.* **7,** 649–666 (1998).
13. Smith, L.J., Daura, X. & van Gunsteren, W.F. Assessing equilibration and convergence in biomolecular simulations. *Proteins* **48,** 487–496 (2002).
14. Snow, C.D., Nguyen, N., Pande, V.S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420,** 102–106 (2002).
15. Zagrovic, B., Snow, C.D., Shirts, M.R. & Pande, V.S. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **323,** 927–937 (2002).
16. Oelschlaeger, P., Schmid, R.D. & Pleiss, J. Modeling domino effects in enzymes: molecular basis of the substrate specificity of the bacterial metallo-beta-lactamases IMP-1 and IMP-6. *Biochemistry* **42,** 8945–8956 (2003).
17. Pang, Y.-P. Three-dimensional model of a substrate-bound SARS chymotrypsin-like cysteine proteinase predicted by multiple molecular dynamics simulations: catalytic efficiency regulated by substrate binding. *Proteins.* **57,** 747–757 (2004).
18. Pornillos, O., Alam, S.L., Davis, D.R. & Sundquist, W.I. Structure of the Tsg101 UEV domain in complex with the PTAP motif of the HIV-1 p6 protein. *Nat. Struct. Biol.* **9,** 812–817 (2002).
19. Chen, H. *et al.* Only one protomer is active in the dimer of SARS 3C-like proteinase. *J. Biol. Chem.* **281,** 13894–13898 (2006).
20. Park, J.G. *et al.* Serotype-selective, small-molecule inhibitors of the zinc endopeptidase of botulinum neurotoxin serotype A. *Bioorg. Med. Chem.* **14,** 395–408 (2006).