MINIREVIEW / MINISYNTHÈSE

Ribosomal synthesis of nonstandard peptides[1]

Taek Jin Kang and Hiroaki Suga

Abstract: It is well known that standard peptides, which comprise proteinogenic amino acids, can act as specific chemical probes to target proteins with high affinity. Despite this fact, a number of peptide drug leads have been abandoned because of their poor cell permeability and protease instability. On the other hand, nonstandard peptides isolated as natural products often exhibit remarkable pharmaco-behavior and stability in vivo. Although it is likely that numerous nonstandard therapeutic peptides capable of recognizing various targets could have been synthesized, enzymes for nonribosomal peptide syntheses are complex; therefore, it is difficult to engineer such modular enzymes to build nonstandard peptide libraries. Here we describe an emerging technology for the synthesis of nonstandard peptides that employs an integrated system of reconstituted cell-free translation and flexizymes. We summarize the historical background of this technology and discuss its current and future applications to the synthesis of nonstandard peptides and drug discovery.

Key words: nonstandard peptide, misaminoacylation, therapeutic peptide, translation.

Résumé : Il est bien connu que les peptides standards, qui comprennent les acides aminés protéinogènes, peuvent agir comme sondes chimiques spécifiques pour cibler des protéines avec une haute affinité. Malgré cela, plusieurs prototypes de drogues peptidiques ont été abandonnés à cause de leur faible perméabilité cellulaire et à leur sensibilité vis-à-vis les protéases. Cependant, des peptides non standards isolés comme produits naturels font souvent preuve de qualités pharmacologiques uniques et de stabilité *in vivo*. Quoiqu'il soit probable que plusieurs peptides thérapeutiques non standards capables de reconnaître différentes cibles puissent être synthétisés, les enzymes de synthèse de peptides non ribosomiques sont complexes; il est alors difficile de concevoir de telles enzymes modulaires pour construire des banques de peptides non standards. Dans cet article, nous décrivons une technologie en émergence qui permet de synthétiser des peptides non standards à l'aide d'un système intégré de traduction acellulaire reconstituée et de flexizymes. Nous résumons ici la base historique de cette technologie et ses applications actuelles et futures pour la synthèse de peptides non standards et pour la découverte de médicaments.

Mots-clés : peptide non standard, mis-aminoacylation, peptide thérapeutique, traduction.

[Traduit par la Rédaction]

Introduction: standard peptides versus nonstandard peptides

The recent successful development of therapeutic proteins has made a significant impact on the pharmaceutical industry (Leader et al. 2008), yet most drug-development initiatives focus on small organic molecules, because of their target flexibility (intra- and extra-cellular target proteins), stability, and oral bioavailability. Despite such advantages, the binding of small organic molecules to proteins mainly relies on the existence of the narrow and deep binding pockets of their target proteins, where they can fit in. Unfortunately, not all therapeutically relevant target proteins have such a characteristic site. Instead, it is known that protein–protein interactions often occur on a wide and shallow protein interface, of which the standard size spans about 1600 Å^2 (Lo Conte et al. 1999). Additionally, small, deep cavities that can serve as binding sites for small organic molecules are rarely found at the interface of protein–protein interaction pairs (Arkin and Wells 2004). Therefore, the development of small organic molecules capable of disrupting protein–protein interactions may be intrinsically difficult. In this sense, therapeutic antibodies and antibody-like proteins are ideal molecules to disrupt these interactions (Leader et al. 2008); however, they are costly, and their poor membrane permeability restricts their use to extracellular targets (Arkin and Wells 2004). Thus, it is desirable to develop therapeutic agents that are big enough to cover the interface of protein targets but still able to penetrate into cells so that the agents can function against both extracellular and intracellular targets.

Received 22 October 2007. Revision received 25 January 2008. Accepted 28 January 2008. Published on the NRC Research Press Web site at bcb.nrc.ca on 20 March 2008.

T.J. Kang and H. Suga.[2] Research Center for Advanced Science and Technology, University of Tokyo, 153-8904 Tokyo; and Department of Chemistry and Biotechnology, Graduate School of Engineering, University of Tokyo, 113-8656 Tokyo, Japan.

[1]This paper is one of a selection of papers published in this Special Issue, entitled CSBMCB — Systems and Chemical Biology, and has undergone the Journal's usual peer review process.

[2]Corresponding author (e-mail: hsuga@rcast.u-tokyo.ac.jp).

 doi:10.1139/O08-009

In the last few decades, a number of short, naturally occurring peptides possessing a variety of biological activities have been discovered. Potencies of these peptides are remarkably high, often exhibiting activities at low concentrations. For instance, human urotensin-II (hU-II), consisting of 11 amino acids, is a potent vasoconstrictor that strongly binds to one of the G-protein-coupled receptors and induces calcium mobilization at subnanomolar concentrations (Ames et al. 1999; Coulouarn et al. 1998). Moreover, peptide fragments artificially designed from substrates of target proteins or their receptors have been shown to retain their biological activities. For instance, a short peptide derived from the erythropoietin receptor (EPO-R) was able to bind to EPO-R and activate the signaling pathway (Naranda et al. 1999). Furthermore, phage display technology has enabled us to screen artificial peptides from random peptide libraries (Parmley and Smith 1988). Along the same lines, phage display selection against EPO-R has given rise to a peptide sequence that has no sequence homology to EPO yet exhibits activity both in vitro and in vivo (Johnson and Jolliffe 2000; Wrighton et al. 1996). Even though the phage display method has been successfully used to generate a variety of peptide sequences that bind to target proteins, such peptides rarely exhibit high therapeutic potencies in vivo. This is because peptides consisting of the 20 proteinogenic (standard) amino acids are generally susceptible to proteases and are often digested before exhibiting their expected biological activities. Thus, to devise protease-resistant peptides based on the phage-selected peptides, each of the analogs must be chemically synthesized and nonproteinogenic amino acids incorporated, follwed by tedious rescreening against the target to optimize such sequences.

Nature overcomes this limitation by synthesizing nonstandard peptides containing unusual monomers, e.g., N^{α}-methylated amino acids, hydroxy acids, and amino acids with D-configuration or non-proteinogenic side-chains. Whereas standard peptides are synthesized by the mRNA-directed polymerization of amino acids by the translation machinery, nonstandard peptides are generally synthesized by the template-independent synthesis machinery consisting of clusters of modular protein enzymes, called nonribosomal peptide sysnthetases (NRPSs) (Fischbach and Walsh 2006). Remarkably, these nonstandard peptides, which had been isolated as natural products, exhibit a wide range of biological activities not only against microorganisms but also in human cells (Schwarzer et al. 2003). Engineering appropriate modules in NRPSs is expected to generate new machineries capable of synthesizing novel kinds of nonstandard peptides. In a preliminary study, a small library of nonstandard peptides was prepared by NRPSs containing engineered donor and acceptor communication domains (Hahn and Stachelhaus 2006). However, current knowledge about the generality and portability of communication domains is yet insufficient to generate randomly shuffled enzyme domains. Thus, it is still a demanding task to further engineer the clusters that produce a variety of nonstandard peptide libraries for the discovery of novel therapeutic molecules.

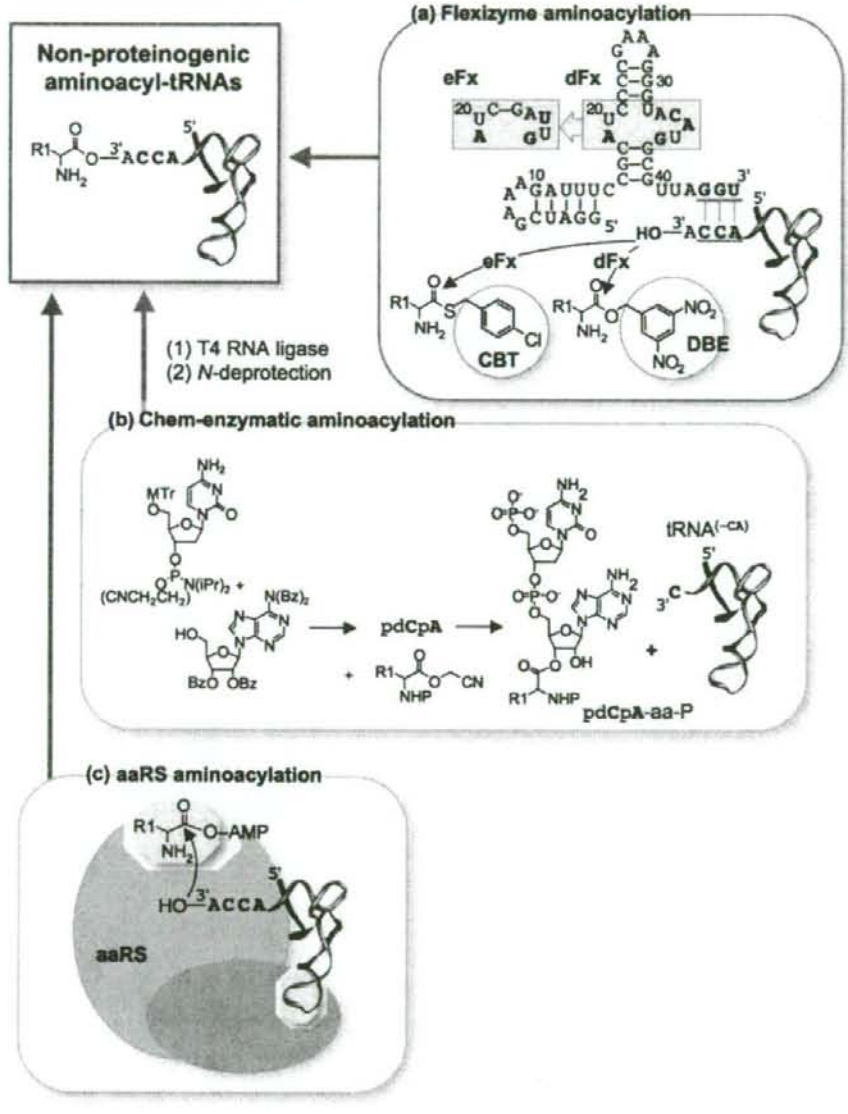
Chemical synthesis has been the only alternative method to generate nonstandard peptide libraries, but it is not necessarily ideal for handling diverse libraries against various protein targets. When the library size reaches one million, it is impractical to test all the individual peptides for their binding properties (Weiss and Chamberlin 2003); thus, diverse libraries are usually subject to various selection strategies to reduce the number of selectants (Turk and Cantley 2003). Additionally, selectants are not usually amplifiable in the chemically synthesized libraries; thus, the absolute amount necessary to identify positive hits restricts practical diversity. Instead, a new technology, termed genetic code reprogramming, has recently been devised for the synthesis of nonstandard peptides in a template-directed format using a reconstituted cell-free translation system. Here, we discuss the development of this emerging technology, as well as its strategies for the synthesis of nonstandard peptides and their libraries and future applications.

Nonsense and 4-base codon suppressions versus genetic code reprogramming

Although the ordinary translation system strictly incorporates 20 proteinogenic amino acids into the nascent peptide chain, some organisms use the same codon for different purposes. For example, UGA and UAG codons occasionally encode selelnocysteine and pyrrolysine, respectively, in a variety of organisms, whereas these codons denote the translation stop signal in general (Zhang and Gladyshev 2007). It is also known that an appropriate manipulation of the translation system enables us to incorporate nonproteinogenic amino acids into peptides. A classical example is that chemically generated misaminoacylated tRNA (alanyl-$tRNA^{Cys}$) supported translation similar to normal enzymatically prepared aminoacyl-tRNAs in vitro (Chapeville et al. 1962; Ehrenstein et al. 1963), indicating that the tRNA aminoacylation event by aminoacyl-tRNA synthetases (aaRSs), and not the ribosome, is the major player controlling translation fidelity. This created the possibility of using the ribosome for making peptides containing nonproteinogenic amino acid(s), and later efforts were devoted to devising methods for assigning nonproteinogenic amino acids to a single codon by using appropriate misacylation technologies.

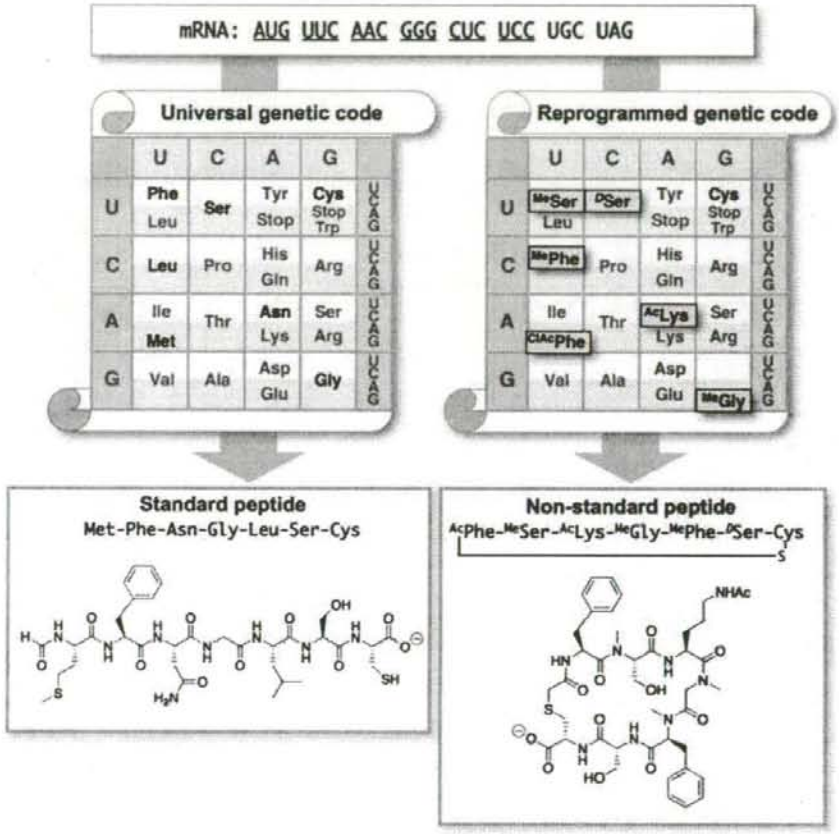
There are 64 possible combinations of triplet nucleotides (codons), all of which are used to code for the 20 amino acids, with the exception of 3 codons, which denote the translation stop signal. These stop codons (or nonsense codons) can be re-assigned to nonproteinogenic amino acid(s) (Noren et al. 1989). However, this method, generally called nonsense suppression, suffers from inherent competition with release factor(s), occasionally yielding low incorporation efficiency, depending upon downstream or upstream codon sets. In addition, this method is suitable only for the incorporation of a single type of nonproteinogenic amino acid into a peptide chain at the specific site assigned by one of the stop codons (usually a UAG stop codon), which is not diverse enough for the construction of a nonstandard peptide library. Complementary to this nonsense suppression, nucleotide quadruplet codons (4-base codons) have been used to assign non-proteinogenic amino acids (Hohsaka et al. 1996), where the programmed frameshift occurs at the 4-base codon as a correct reading frame. This method enabled the incorporation of two, and occasionally three, nonproteinogenic amino acids charged onto

Fig. 1. Preparation of misaminoacyl-tRNAs. (*a*) Flexizyme method (left). Two ribozymes, dFx and eFx, show different substrate specificities, as shown. Flexizyme recognizes only the 3 nucleotides at the 3′-end of tRNA (N_{73}-C_{75}) and the active ester of amino acid substrates, thus this method is compatible with a wide range of acids and tRNAs. (*b*) Chem-enzymatic method (top). The terminal dinucleotide (pdCpA) is chemically synthesized using the standard phosphoramidite method (1); the substrate amino acid (aa) protected by an appropriate protective group is chemically linked to the 3′ end of pdCpA (2); the resulting pdCpA-aa is subsequently ligated to the truncated tRNA (tRNA–3′CA) lacking the 3′-terminal CA by the action of T4 RNA ligase (3); deprotection of the protective group to generate aa-tRNA (4). MTr, methoxytrityl; Bz, benzoyl. (*c*) Enzymatic method using engineered aminoacyl-tRNA synthetase (aaRS, bottom). Both the rational design and the molecular evolution were reported to render the extremely high specificity of some aaRSs. Unlike the previous two methods, this approach enables the large scale syntheses of nonstandard amino acids by expressing engineered aaRS in the host cell. However, because of high specificity, each engineered aaRS must be prepared for each nonstandard amino acid and for each anticodon. So far, the choice of nonstandard amino acids is limited and the anticodon is also limited to the anti-UAG codon.



tRNAs bearing 4-base anticodons into the peptide chain (Ohtsuki et al. 2005). However, because 4-base codons must be designed based on rarely used codons in *Escherichia coli*, the number of usable codons is still restricted, particularly upon using exotic nonproteinogenic acids that are often difficult to incorporate into a peptide chain, and this method suffers from incomplete synthesis of the peptide as a result of undesired reading of the 4-base codon(s) by com-

Fig. 2. Genetic reprogramming approach for the preparation of nonstandard peptides. When the mRNA sequence shown is translated to a peptide, the translation system faithfully follows the universal code book (shown on the left) in the usual in vitro translation system (PURE system). The resulting heptapeptide (fMet-Phe-Asn-Gly-Leu-Ser-Cys) is shown below. In this example, by removing Met, Phe, Asn, Gly, Leu, and Ser (and (or) corresponding aaRSs) from the translation system (wPURE system), codons for those amino acids can be re-assigned to nonstandard amino acids. For this genetic reprogramming, the key component is misaminoacyl-tRNA (^{ClAc}Phe-tRNA$_{ini}$, ^{Me}Ser-tRNA$_{GAA}$, ^{Ac}Lys-tRNA$_{GUU}$, ^{Me}Gly-tRNA$_{CCC}$, ^{Me}Phe-tRNA$_{GAG}$, and ^{D}Ser-tRNA$_{GGA}$, in this example; ^{ClAc}Phe, N^{α}-(2-chloroacetyl)phenylalanine; ^{Ac}Lys, N^{ε}-acetyl-lysine; ^{Me}Gly and ^{Me}Phe, N^{α}-methyl-glycine and N^{α}-methyl-phenylalanine, repectively). The resulting nonstandard peptide is shown below.

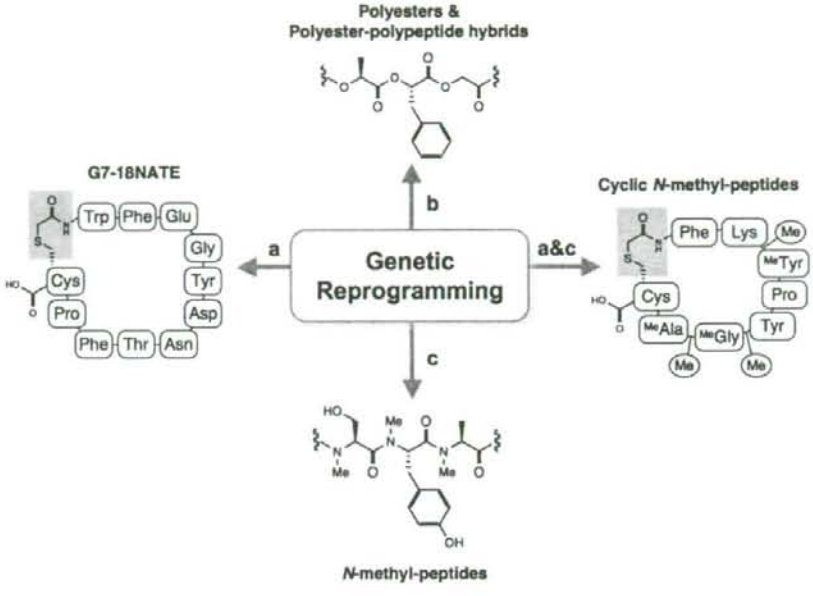


peting aminoacyl-tRNA(s) bearing the normal 3-base anticodon(s).

In addition to the above limitations, the step of tRNA aminoacylation with nonproteinogenic amino acids relies on laborious and technically demanding chem-enzymatic processes (Hecht et al. 1978; Heckler et al. 1984; Noren et al. 1989; Robertson et al. 1989). This technical barrier made this method so specialized that only a subset of researchers was able to use it (Fig. 1). More recently, Schultz and other groups have succeeded in developing mutant aaRSs that charge a certain family of nonproteinogenic amino acids (mostly aromatic amino acid analogs) onto orthogonal tRNA (Kiga et al. 2002; Wang and Schultz 2005; Yoo and Tirrell 2007; Zhang et al. 2004). However, this method is limited to the incorporation of a single nonproteinogenic amino acid, and also the choice of the amino acids remains limited (Fig. 1).

More recently, independent efforts were made to overcome such limitations in conventional nonproteinogenic amino acid mutagenesis by Foster et al., followed by those of several other groups (Forster et al. 2003; Josephson et al. 2005; Murakami et al. 2006). In contrast to the aforementioned methods involving nonsense or 4-base codon suppression, this method involves the reassignment of arbitral codons from proteinogenic to nonproteinogenic amino acids. However, nucleotide triplet codons in the mRNA pass genetic information to the peptide through the tightly regulated aminoacylation of cognate tRNAs, making it very difficult to reassign the genetic code in the usual translation system. A key technology that makes it possible to break the tight regulation of the correspondence is a reconstituted *E. coli* cell-free translation system, often referred to as PURE (protein synthesis using recombinant elements) system (Shimizu et al. 2001). The most important feature of this translation

Fig. 3. Genetic code reprogramming and its applications. Genetic code reprogramming made it possible to incorporate the key amino acids into a peptide for various purposes. (*a*) N^{α}-(2-chloroacetyl)tyrosine was incorporated into the sequence of G7-18NATE via initiation codon reprogramming. Subsequent cyclization through thioether yielded a cyclized G7-18NATE analog. (*b*) α-Hydroxy acids, lactic acid, phenyllactic acid, and glycolic acid, were incorporated into a peptide structure consecutively, yielding an ester-backbone-containing peptide. (*c*) N^{α}-Methyl-serine, N^{α}-methyl-phenylalanine, and N^{α}-methyl-alanine are incorporated into a peptide. The resulting peptide has a nonstandard backbone that can be found in many therapeutically important natural products. This approach can be combined with the initiation codon reprogramming to ribosomally synthesize cyclized N^{α}-methylated peptides, yielding a structure that has great potential in peptide therapeutics.



system is that certain amino acids and aaRSs can be withdrawn from the translation elements. For example, when the translation system is deprived of leucine and its corresponding aaRS, leucinyl-tRNA cannot be synthesized in the system, and thus is practically withdrawn, even though corresponding tRNA is present in the system. Using such a withdrawn PURE system, termed wPURE, vacant codons can be created at a researcher's will (Fig. 2). The significant differences between this method and the nonsense or 4-base codon suppression are (*i*) because natural aminoacyl-tRNAs are removed from the translation system, competition between natural and nonstandard aminoacyl-tRNAs would not occur, and (*ii*) through appropriate selection of codons for the nonstandard amino acid, the maximal sense suppression efficiency can be obtained. A major barrier that remains for this method is how nonproteinogenic amino acids can be assigned to these vacant codons, i.e., how can they be attached onto the desired tRNAs capable of reading the vacant codons. Clearly, a nonlaborious and reliable method is necessary to perform this process.

Methods for tRNA aminoacylation, the key technology

The Szostak group showed a pioneering example of codon table redesigning (Josephson et al. 2005). Using nonproteinogenic amino acids compatible with aaRSs, they reassigned the universal genetic code to one that contains 12 nonproteinogenic amino acids, and also showed that this translation system was compatible with the newly assigned genetic code. The translation system resulted in peptides containing nonproteinogenic amino acids, as designated by the mRNA sequence. This technology, together with their proprietary technology, mRNA display, had opened a possibility for in vitro screening of nonstandard peptide aptamers (peptides capable of binding to targets). However, since this system utilizes aaRSs for mischarging tRNAs with nonproteinogenic amino acids, only those that structurally resemble proteinogenic amino acids (and thus, are compatible with aaRSs) can be used, thereby limiting the variety of peptides that can be used. A more serious problem of this approach is contamination of proteinogenic amino acids carried over into the wPURE system with purified ribosome or recombinant factors. When exotic nonproteinogenic amino acids, which are usually inefficiently incorporated into the nascent peptide chain are used for the suppressions, a small amount of such unavoidable proteinogenic contaminants readily outcompete the nonproteinogenic amino acids, leading to peptides composed primarily of these standard amino acids.

Obviously, a general method for the preparation of nonstandard aminoacyl-tRNAs is a prerequisite for genetic reprogramming. Although the chem-enzymatic method is suitable to this application, and indeed, was used in the earlier work by Foster et al. (Forster et al. 2003), its technical difficulties and laboriousness hinder its widespread use in creating peptide libraries with diverse kinds of non-proteino-

Table 1. Yields of acyl-tRNAs.

Acid substrate	Flexizyme	Yield (%)
Standard amino acids		
Ala-DBE	dFx	36[a]
Asn-DBE	dFx	22[b]
Asp-DBE	dFx	52[b]
Cys-DBE	dFx	46[b]
Gln-DBE	dFx	46[b]
Glu-DBE	dFx	17[a]
Gly-DBE	dFx	39[a]
His-DBE	dFx	29[a]
Leu-DBE	dFx	37[a]
Lys-DBE	dFx	36[a]
Met-DBE	dFx	35[a]
Phe-CME	eFx	47[a]
Pro-DBE	dFx	37[b]
Ser-DBE	dFx	38[a]
Trp-CME	eFx	36[a]
Tyr-CME	eFx	34[a]
Val-DBE	dFx	13[a]
Amino acids with nonstandard side-chains		
Aly-DBE	dFx	33[a]
Bly-DBE	dFx	30[a]
Cit-DBE	dFx	35[a]
α-Hydroxy acids		
Hbi-DBE	dFx	25[a]
Hle-DBE	dFx	51[b]
Hph-CME	eFx	82[b]
β-Amino acid and N^{α}-methyl amino acid		
Bal-DBE	dFx	17[a]
Mle-DBE	dFx	55[b]

Note: Yields were calculated using either a streptavidin-dependent gel-shift assay or acid PAGE, depending on the substrate (Murakami et al. 2006). Amino acids are represented in standard three-letter abbreviations except as follows: Aly, ε-*N*-acetyl-lysine; Bly, ε-*N*-biotinyl-lysine; Cit, L-citrulline; Hbi, δ-*N*-biotinyl-(*S*)-hydroxybutanoic acid; Hle, (*S*)-3-isoproyllactic acid; Hph, (*S*)-3-phenyllactic acid; Bal, β-alanine; Mle, α-*N*-methyl-leucine. Substrates were activated by either 3,5-dinitrobenzyl (DBE) or cyanomethyl ester (CME).

[a]Yield calculated using a streptavidin-dependent gel-shift assay.

[b]Yield calculated using acid PAGE.

genic amino acids. Instead, an artificial RNA enzyme (ribozyme) capable of catalyzing aminoacylation of tRNAs was generated from a random sequence pool of RNA by our group (Saito et al. 2001), and later, this ribozyme turned into a highly flexible tool for tRNA aminoacylation and was named the flexizyme system (Murakami et al. 2006). The system consists of two kinds of flexizymes, called dFx and eFx, which can be used depending upon the choice of a leaving group on the substrates; dFx for 3,5-dinitobenzyl ester (DBE), and eFx for 4-chlorobenzyl thioester (CBT) or cyanomethyl ester (CME) (Fig. 1). Because the flexizymes recognize their cognate leaving group, and not the side-chain or the free amino group of a given substrate, they are able to charge a wide variety of α-amino acids with nonproteinogenic side-chains. However, the most remarkable feature is that the combination of these two flexizymes gives virtually no restriction of substrates such as N^{α}-methylated amino acids, D-α-amino acids, β-amino acids, and even α-hydroxy acids (Table 1). Moreover, flexizymes recognize N_{73}–C_{75} of tRNA by base pairings with G43–U45, where C_{74} and C_{75} are common to all tRNAs and N_{73} can be any of A, G, and U (even C can be accepted if the incubation time is prolonged); thus virtually any tRNA can be a substrate for flexizymes (Fig. 1).

Thus, we combined the flexizyme system with a wPURE system to demonstrate the genetic code reprogramming for the synthesis of nonstandard peptides. In the first demonstration of the combined systems, three codons (AGU, AAC, and CAG for serine, asparagines, and glutamine, respectively) were reassigned to three nonstandard amino acids (ε-*N*-acetyl-lysine, citrulline, and *p*-iodo-phenylalnine, respectively) so that a 17-mer peptide possessing 6 nonproteinogenic amino acids could be synthesized by the ribosome as efficiently as the original sequence with standard amino acids (Murakami et al. 2006). Typical yields range from several to 20 μmol/L in the translation mixture. With this foundation, our efforts have been made to prepare a wide variety of nonstandard peptides with highly altered chemical structures. We will describe some of our most recent results to exemplify the versatility of this new method.

Ribosomal synthesis of nonstandard polypeptides and polyesters

The SH2 domain of Grb7 plays a role in signaling by binding to an intracellular phosphorylated tyrosine of several receptor tyrosine kinases. Pero et al. performed a selection of peptides against the SH2 domain using phage display and isolated novel high affinity peptides (Pero et al. 2002). G7-18, one of the most potent peptides selected, has the characteristics of a nonphosphorylated and cyclic form, closed by a disulfide bond between two internal cysteine residues. Unfortunately, this disulfide bond in G7-18 was reducible so that its linear form not only failed to exhibit activity but also was susceptible to proteases in vivo. The same team ingeniously substituted the disulfide bond with a thioether bond by chemical synthesis; the N terminus of the cysteine residue of G7-18 was substituted with a 2-chloroacetyl group (ClAc), resulting in the intramolecular attack of the C-terminal cysteine side-chain to the 2-position of the acetyl group. Despite the fact that this new peptide, called G7-18NATE, lost approximately 10-fold of its affinity to Grb7 compared with that of G7-18, it still exhibited antitumor activity in an animal study. We recently succeeded in the ribosomal synthesis of G7-18NATE using initiation codon reprogramming, in which ^{ClAc}Trp was used to prime the translation in a methionine-depleted wPURE system (Goto et al. 2008). Initiator tRNA was charged with ^{ClAc}Trp by the action of flexizyme, and the resulting aminoacyl-tRNA was used for the translation initiation instead of initiator tRNA charged with α-*N*-formyl-methione. Remarkably, the peptide was spontaneously and only intramolecularly cyclized through the thioether bond upon the completion of translation, and therefore no extra treatment was necessary. We envision that our method will allow us to re-investigate the sequence of G7-18NATE by coupling initiation codon reprogramming with an appropriate in vitro display system.

Since this genetic code reprogramming system lacks competitors for the incorporations of nonproteinogenic amino acids, it is possible to perform multiple incorporations or the polymerization of monomers that would alter the backbone from the ordinary peptide bond to other types of bonds. For instance, we have recently performed mRNA-directed polyester synthesis using seven varieties of α-hydroxy acids that were each individually assigned to one of seven codons (Ohta et al. 2007). This work represents the first demonstration of up to 12 consecutive additions of α-hydroxy acids with various different compositions designated by codons on mRNA. We also recently performed the mRNA-directed synthesis of linear and cyclic *N*-methyl-peptides (Kawakami et al. 2008). As a proof-of-concept experiment, up to 10 successive incorporations of N^{α}-methyated amino acids with six different kinds of side-chains were performed. In this work, we also combined this method with the aforementioned thioether cyclization method using the ClAc group, showing that cyclic *N*-methyl-peptides with three or four *N*-methylated backbone residues could be synthesized. This work has opened the possibility of the mRNA-encoded synthesis of *N*-methyl-peptide libraries for the screening of biologically active molecules.

Conclusion and perspectives

Since the advent of the combinatorial synthesis of peptides, peptides have drawn significant attention with the anticipation of finding novel drugs. However, only a few therapeutic peptides have reached the market because of problems associated with poor stability and cell-membrane permeability. A lesson from peptide-like natural products has given a possible direction; the peptide should have non-standard structures with not only side-chains but a cyclic and N^{α}-methylated backbone, as well. We now have a new tool for synthesizing such nonstandard peptides in a cell-free translation system so that their sequences can be encoded by oligonucleotides (mRNA or cDNA) that are readily amplifiable and sequence readable by conventional molecular biological techniques.

A controllable translation system and a flexible tool for the preparation of a wide variety of nonstandard aminoacyl-tRNAs are essential to genetic code reprogramming. As described here, the combined use of wPURE system and the flexizyme technology is a reliable methodology for the ribosomal syntheses of nonstandard peptides containing the desired modifications (Fig. 3). By means of genetic code reprogramming, we can expect a whole new era of peptide chemistry that is compatible not only with cell-based high-throughput screening, but with in vitro selection methods, such as mRNA or ribosome display against isolated targets, as well. We are on the verge of witnessing the discovery of active nonstandard peptides against various therapeutic targets.

Acknowledgements

This work was supported by grants from the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (S) (16101007) and the US National Institutes of Health (GM59159).

References

Ames, R.S., Sarau, H.M., Chambers, J.K., Willette, R.N., Alyar, N.V., Romanic, A.M., et al. 1999. Human urotensin-II is a potent vasoconstrictor and agonist for the orphan receptor GPR14. Nature, **401**: 282–286. doi:10.1038/45809. PMID:10499587.

Arkin, M.R., and Wells, J.A. 2004. Small-molecule inhibitors of protein–protein interactions: Progressing towards the dream. Nat. Rev. Drug Discov. **3**: 301–317. doi:10.1038/nrd1343. PMID:15060526.

Chapeville, F., Ehrenstein, G.V., Benzer, S., Weisblum, B., Ray, W.J., and Lipmann, F. 1962. On role of soluble ribonucleic acid in coding for amino acids. Proc. Natl. Acad. Sci. U.S.A. **48**: 1086–1092. doi:10.1073/pnas.48.6.1086. PMID:13878159.

Coulouarn, Y., Lihrmann, I., Jegou, S., Anouar, Y., Tostivint, H., Beauvillain, J.C., et al. 1998. Cloning of the cDNA encoding the urotensin II precursor in frog and human reveals intense expression of the urotensin II gene in motoneurons of the spinal cord. Proc. Natl. Acad. Sci. U.S.A. **95**: 15803–15808. doi:10.1073/pnas.95.26.15803. PMID:9861051.

Ehrenstein, G., Weisblum, B., and Benzen, S. 1963. Function of sRNA as amino acid adaptor in synthesis of hemoglobin. Proc. Natl. Acad. Sci. U.S.A. **49**: 669–675. doi:10.1073/pnas.49.5.669. PMID:16591086.

Fischbach, M.A., and Walsh, C.T. 2006. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: Logic, machinery, and mechanisms. Chem. Rev. **106**: 3468–3496. doi:10.1021/cr0503097. PMID:16895337.

Forster, A.C., Tan, Z.P., Nalam, M.N.L., Lin, H.N., Qu, H., Cornish, V.W., and Blacklow, S.C. 2003. Programming peptidomimetic syntheses by translating genetic codes designed de novo. Proc. Natl. Acad. Sci. U.S.A. **100**: 6353–6357. doi:10.1073/pnas.1132122100. PMID:12754376.

Goto, Y., Ohta, A., Sako, Y., Yamagishi, Y., Murakami, H., and Suga, H. 2008. Reprogramming the translation initiation for the synthesis of physiologically stable cyclic peptides. ACS Chem. Biol. **3**: 120–129. doi:10.1021/cb700233t. PMID:18215017.

Hahn, M., and Stachelhaus, T. 2006. Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. Proc. Natl. Acad. Sci. U.S.A. **103**: 275–280. doi:10.1073/pnas.0508409103. PMID:16407157.

Hecht, S.M., Alford, B.L., Kuroda, Y., and Kitano, S. 1978. "Chemical aminoacylation" of tRNA's. J. Biol. Chem. **253**: 4517–4520. PMID:248056.

Heckler, T.G., Chang, L.H., Zama, Y., Naka, T., Chorghade, M.S., and Hecht, S.M. 1984. T4 RNA ligase mediated preparation of novel "chemically misacylated" tRNAPhes. Biochemistry, **23**: 1468–1473. doi:10.1021/bi00302a020. PMID:6372858.

Hohsaka, T., Ashizuka, Y., Murakami, H., and Sisido, M. 1996. Incorporation of nonnatural amino acids into streptavidin through in vitro frame-shift suppression. J. Am. Chem. Soc. **118**: 9778–9779. doi:10.1021/ja9614225.

Johnson, D.L., and Jolliffe, L.K. 2000. Erythropoietin mimetic peptides and the future. Nephrol. Dial. Transplant. **15**: 1274–1277. doi:10.1093/ndt/15.9.1274. PMID:10978375.

Josephson, K., Hartman, M.C.T., and Szostak, J.W. 2005. Ribosomal synthesis of unnatural peptides. J. Am. Chem. Soc. **127**: 11727–11735. doi:10.1021/ja0515809. PMID:16104750.

Kawakami, T., Murakami, H., and Suga, H. 2008. Messenger RNA-programmed incorporation of multiple *N*-methyl-amino acids into linear and cyclic peptides. Chem. Biol. **15**: 32–42. doi:10.1016/j.chembiol.2007.12.008. PMID:18215771.

Kiga, D., Sakamoto, K., Kodama, K., Kigawa, T., Matsuda, T., Yabuki, T., et al. 2002. An engineered *Escherichia coli* tyrosyl-

tRNA synthetase for site-specific incorporation of an unnatural amino acid into proteins in eukaryotic translation and its application in a wheat germ cell-free system. Proc. Natl. Acad. Sci. U.S.A. **99**: 9715–9720. doi:10.1073/pnas.142220099. PMID:12097643.

Leader, B., Baca, Q.J., and Golan, D.E. 2008. Protein therapeutics: a summary and pharmacological classification. Nat. Rev. Drug Discov. **7**: 21–39. doi:10.1038/nrd2399. PMID:18097458.

Lo Conte, L., Chothia, C., and Janin, J. 1999. The atomic structure of protein–protein recognition sites. J. Mol. Biol. **285**: 2177–2198. doi:10.1006/jmbi.1998.2439. PMID:9925793.

Murakami, H., Ohta, A., Ashigai, H., and Suga, H. 2006. A highly flexible tRNA acylation method for non-natural polypeptide synthesis. Nat. Methods, **3**: 357–359. doi:10.1038/nmeth877. PMID:16628205.

Naranda, T., Wong, K., Kaufman, R.I., Goldstein, A., and Olsson, L. 1999. Activation of erythropoietin receptor in the absence of hormone by a peptide that binds to a domain different from the hormone binding site. Proc. Natl. Acad. Sci. U.S.A. **96**: 7569–7574. doi:10.1073/pnas.96.13.7569. PMID:10377456.

Noren, C.J., Anthonycahill, S.J., Griffith, M.C., and Schultz, P.G. 1989. A general-method for site-specific incorporation of unnatural amino-acids into proteins. Science, **244**: 182–188. doi:10.1126/science.2649980. PMID:2649980.

Ohta, A., Murakami, H., Higashimura, E., and Suga, H. 2007. Synthesis of polyester by means of genetic code reprogramming. Chem. Biol. **14**: 1315–1322. PMID:18096500.

Ohtsuki, T., Manabe, T., and Sisido, M. 2005. Multiple incorporation of non-natural amino acids into a single protein using tRNAs with non-standard structures. FEBS Lett. **579**: 6769–6774. doi:10.1016/j.febslet.2005.11.010. PMID:16310775.

Parmley, S.F., and Smith, G.P. 1988. Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. Gene, **73**: 305–318. doi:10.1016/0378-1119(88)90495-7. PMID:3149606.

Pero, S.C., Oligino, L., Daly, R.J., Soden, A.L., Liu, C., Roller, P.P., et al. 2002. Identification of novel non-phosphorylated ligands, which bind selectively to the SH2 domain of Grb7. J. Biol. Chem. **277**: 11918–11926. doi:10.1074/jbc.M111816200. PMID:11809769.

Robertson, S.A., Noren, C.J., Anthonycahill, S.J., Griffith, M.C., and Schultz, P.G. 1989. The use of 5′-phospho-2 deoxyribocytidylylriboadenosine as a facile route to chemical aminoacylation of transfer-RNA. Nucleic Acids Res. **17**: 9649–9660. doi:10.1093/nar/17.23.9649. PMID:2602139.

Saito, H., Kourouklis, D., and Suga, H. 2001. An in vitro evolved precursor tRNA with aminoacylation activity. EMBO J. **20**: 1797–1806. doi:10.1093/emboj/20.7.1797. PMID:11285242.

Schwarzer, D., Finking, R., and Marahiel, M.A. 2003. Nonribosomal peptides: from genes to products. Nat. Prod. Res. **20**: 275–287. doi:10.1039/b111145k.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. 2001. Cell-free translation reconstituted with purified components. Nat. Biotechnol. **19**: 751–755. doi:10.1038/90802. PMID:11479568.

Turk, B.E., and Cantley, L.C. 2003. Peptide libraries: at the crossroads of proteomics and bioinformatics. Curr. Opin. Chem. Biol. **7**: 84–90. doi:10.1016/S1367-5931(02)00004-2. PMID:12547431.

Wang, L., and Schultz, P.G. 2005. Expanding the genetic code. Angew. Chem. Int. Ed. **44**: 34–66. doi:10.1002/anie.200460627.

Weiss, G.A., and Chamberlin, R. 2003. Bridging the synthetic and biopolymer worlds with peptide-drug conjugates. Chem. Biol. **10**: 201–202. doi:10.1016/S1074-5521(03)00056-5. PMID:12670531.

Wrighton, N.C., Farrell, F.X., Chang, R., Kashyap, A.K., Barbone, F.P., Mulcahy, L.S., et al. 1996. Small peptides as potent mimetics of the protein hormone erythropoietin. Science, **273**: 458–463. doi:10.1126/science.273.5274.458. PMID:8662529.

Yoo, T.H., and Tirrell, D.A. 2007. High-throughput screening for methionyl-tRNA synthetases that enable residue-specific incorporation of noncanonical amino acids into recombinant proteins in bacterial cells. Angew. Chem. Int. Ed. **46**: 5340–5343. doi:10.1002/anie.200700779.

Zhang, Y., and Gladyshev, V.N. 2007. High content of proteins containing 21st and 22nd amino acids, selenocysteine and pyrrolysine, in a symbiotic deltaproteobacterium of gutless worm Olavius algarvensis. Nucleic Acids Res. **35**: 4952–4963. doi:10.1093/nar/gkm514. PMID:17626042.

Zhang, Z., Alfonta, L., Tian, F., Bursulaya, B., Uryu, S., King, D.S., and Schultz, P.G. 2004. Selective incorporation of 5-hydroxytryptophan into proteins in mammalian cells. Proc. Natl. Acad. Sci. U.S.A. **101**: 8882–8887. doi:10.1073/pnas.0307029101. PMID:15187228.

Highlighted paper selected by Editor-in-chief

Determination of Ligand-Binding Sites on Proteins Using Long-Range Hydrophobic Potential

Noriyuki YAMAOTSU,*[,a] Akifumi ODA,[b,1)] and Shuichi HIRONO[a]

[a] *Laboratory of Physical Chemistry for Drug Design, School of Pharmaceutical Sciences, Kitasato University; 5–9–1 Shirokane, Minato-ku, Tokyo 108–8641, Japan: and* [b] *Discovery Laboratories, Toyama Chemical Co., Ltd.; 2–4–1 Shimookui, Toyama 930–8508, Japan.* Received April 7, 2008; accepted May 28, 2008; published online May 30, 2008

Here we developed a new program, HydrophoBicity On a Protein (HBOP), to find the ligand-binding site of a protein using the long-range hydrophobic-potential function estimated from the experimental data of Israelachvili and Pashley. We calculated the hydrophobic-potential energies at each grid point of a lattice around a protein using the potential function. The hydrophobic potential was evaluated using the carbon atoms of the hydrophobic residues, with the exception of those of the amide groups. We tested HBOP on 26 types of protein (72 protein–ligand complexes), the three-dimensional structures of which were determined experimentally. Although only one hydrophobic function was used, HBOP could successfully identify the binding sites in all of the proteins tested. Moreover, in 24 of the proteins, the binding sites were located in the most hydrophobic region. Surprisingly, the binding sites on sugar binding proteins were the most hydrophobic sites. It implies that the hydrophobic interaction plays an important role in the formation of protein–ligand complexes.

Key words binding site; hydrophobic interaction; structure-based drug design; binding pocket; hydrophobicity; ligand

Identifying the ligand-binding site in the three-dimensional (3D) structure of a target protein is a starting point for the rational design of therapeutic agents. Many procedures have been developed for the detection of binding sites in protein 3D structures. These are divided into two types of algorithms: those based on geometrical cavity detection in a protein, and those based on the detection of interaction points on a protein. The first type includes a surface method (MS),[2,3)] grid-surface methods (Cavity Search[4)] and VOIDOO,[5)]) alpha-shape methods derived from the Voronoi diagram (CAST[6)] and VOLBL[7)]), a layer method using a probe (PASS),[8)] and a mapping method using the hydrophobic groups on a protein surface.[9)] These procedures are often unable to detect binding sites close to the protein surface or to identify more open binding sites. The second type includes the detection of interaction points by a probe using grid searching methods (GRID[10—13)], DOCK[14—16)], Q-SiteFinder[17)] and SiteMap[18)]) and a random searching method (MCSS).[19—22)] These procedures used several probes (methane, water, *etc.*) to detect pockets on a protein, and then evaluated likelihoods of the binding site with detected pockets by physical or empirical functions (van der Waals, electrostatic, hydrogen-bonding, *etc.*). However, because the functions of these interactions are short range, the representation of a binding site on a protein is discrete and does not correspond to the whole binding site.

Hydrophobicity is a key parameter for understanding drug activity[23,24)]; however, it has rarely been used in the identification of binding sites, because it is difficult to define. Hydrophobicity is usually measured and reported as a $\log P$ value, where P is the partition coefficient of the molecules in octanol/water. However, the dependency of the hydrophobic interaction on distance has remained unknown, and it has therefore been approximated by various functions of distance. Hydrophobic-potential functions that use lipophilic constants based on $\log P$ are known as molecular lipophilicity potentials (*MLP*s).[25—33)] Audry's first *MLP* was based on a hyperbolic distance function, Eq. 1.[25,26)]

$$MLP=f_i f_j/(1+d_{ij}) \tag{1}$$

Here, f_i and f_j are the lipophilic constants of two interacting fragments, d_{ij} is the distance between the fragments. Furet *et al.* also used a hyperbolic distance function.[27)] Fauchère *et al.* introduced an exponential distance function into MLP, Eq. 2.[28)]

$$MLP=f_i f_j \exp(-d_{ij}) \tag{2}$$

Kellogg *et al.* proposed HINT (Hydrophobic INTeractions) function which is atom-based, Eq. 3.[29—32)]

$$MLP=s_i a_i s_j a_j \exp(-d_{ij}) \tag{3}$$

Here, hydrophobic atom constants, a_i and a_j, are multiplied by the solvent-accessible surface area of each atom, s_i and s_j. Gaillard *et al.* introduced the decay length of 2 Å into an exponential function, Eq. 4.[33)]

$$MLP=f_i f_j \exp(-d_{ij}/2) \tag{4}$$

All of *MLP*s were fitted to the $\log P$ values of compounds on the assumption that they were situated in contact positions or associated positions. Therefore, the short decay length was calculated as 1 or 2 Å (Eqs. 2—4).

Israelachvili and Pashley measured the hydrophobic interaction between two monolayer-coated mica surfaces in aqueous solutions, and approximated the results using the exponential expression shown in Eqs. 5a—c.[34)]

$$\Delta G_H=-CR_{ij}D_0 \exp(-D_{ij}/D_0) \tag{5a}$$

$$R_{ij}=R_i R_j/(R_i+R_j) \tag{5b}$$

$$D_{ij}=d_{ij}-(R_i+R_j) \tag{5c}$$

Here, ΔG_H is the pair hydrophobic interaction free energy. According to this equation, the lower the hydrophobic free energy, the stronger the hydrophobicity. Using a monolayer of the surfactant hexadecyl-trimethyl-ammonium bromide, $C=0.20$ kcal mol^{-1} Å^{-1} and the decay length of $D_0=10$ Å. The D_{ij} is the minimum distance between the surfaces of the spheres, d_{ij} is the distance between the centres of the spheres,

* To whom correspondence should be addressed. e-mail: yamaotsun@pharm.kitasato-u.ac.jp.

and R_i and R_j are the radii of two interacting spheres. The unit of energy is kcal/mol, and the unit of length is Å. The decay length is more than 10 Å (*e.g.*, 175 Å for the another surfactant dimethyl-dioctadecyl-ammonium bromide),[34—43] and is much longer than that of the *MLP*s (1 or 2 Å). Thus, because long-range hydrophobic effects play important roles in macromolecules, Eq. 5 is more appropriate for detecting the ligand-binding sites of proteins.

Here we developed new programs, HydrophoBicity On a Protein (HBOP)/HydrophoBic SITE (HBSITE), in order to identify the binding site in a protein using the only hydrophobic-interaction function proposed by Israelachvili and Pashley.[34] We tested our method on 26 proteins (72 protein–ligand complexes), the 3D structures of which were experimentally determined.

MATERIALS AND METHODS

Construction of a Test Set To test our method, we needed datasets of protein–ligand complexes for which the binding geometry was reliable. We thus decided to use a test set constructed from 26 proteins (72 complexes), which were selected from Eldridge's datasets[44] because they were as different as possible from each other, with the exception of immunoglobulins. The Protein Data Bank (PDB) identification (ID) codes of these complexes are shown in Table 1.

For cathepsin D and thrombin, both the light and the heavy chains were treated, the human immunodeficiency virus-1 (HIV-1) protease was a homodimer, and all other proteins were monomers. Initially, all additional molecules (such as ligands, cofactors, water, and ions) were deleted. The original positions of the missing atoms in the PDB structure and the hydrogen atoms were then generated by SYBYL 6.8.1.[45] Only the missing atoms and the hydrogen atoms were minimized using SYBYL with the convergence criterion for the energy gradient of 0.05 kcal mol^{-1} Å^{-1}. The minimizations used the AMBER 1986 force field and the predefined atomic charges of the amino acids.[46,47] The cut-off distance for the non-bonded interactions was 10 Å. A distance-dependent dielectric constant of 4r was used.

Finding the Binding Site The calculations were performed using our new program, HBOP. The grid points of a lattice were generated around the protein surface (Fig. 1a). The thickness of the lattice was 10 Å, and the grid spacing was 1 Å. The grid points were removed if the probe spheres on them impacted with the protein surface, and then the isolated grid points without adjoining points were eliminated. The hydrophobic free energy at each grid point was estimated using the pair interaction free energy function determined mechanically by Israelachvili and Pashley, Eq. 6.[34]

$$\Delta G_H = -2.0 R_{ij} \exp(-D_{ij}/10) \quad (6a)$$

$$R_{ij} = R_i R_j/(R_i + R_j) \quad (6b)$$

$$D_{ij} = d_{ij} - (R_i + R_j) \quad (6c)$$

R_i and R_j were the radii of the carbon atom of the protein and the probe on the grid point, respectively. The probe was the sp^3 carbon atom, and R_j was equal to 1.52 Å. d_{ij} is the distance between the center of the carbon atom of the protein and the grid point. The hydrophobic potential was calculated using only the carbon atoms, with the exception of the amide carbon, of hydrophobic residues (Gly, Ala, Val, Leu, Ile, Met, Trp, Phe, and Pro). Because the magnitude of the hydrophobic interaction is increased with increasing hydrophobicity on the surface[36,37] and reduced with increasing electrolyte concentration,[37,39,40,43] we used only hydrophobic residues to calculate. No cut-off was used in the calculation of the hydrophobic energy. The radii of the carbon atoms implemented in SYBYL 6.8.1 were used by the program. The grid points were divided into 20 levels by 5% of ($\Delta G_H^{max} - \Delta G_H^{min}$), where ΔG_H^{min} was the minimum hydrophobic energy in the system and ΔG_H^{max} was the maximum. The 20 levels were ranked from the most hydrophobic (designated as HB1) to the least hydrophobic (designated as HB20). The hydrophobic site was determined as the region that consisted of grid points with ΔG_H values that fulfilled the empirically derived condition $\Delta G_H^{min} \leq \Delta G_H \leq (0.7\Delta G_H^{min} + 0.3\Delta G_H^{max})$—that is, grid points with hydrophobic levels ranging from HB1 to HB6 (Fig. 1b).

Using a utility program for grid-clustering, HBSITE, with a clustering radius of 1.1 Å, clusters with volumes larger than or equal to 10 Å^3 (10 grid points) were adopted as candidates for the binding site (Fig. 1c). However, a lower limit for triose phosphate isomerase of 3 Å^3 was used, because the ligand was relatively small.

HBOP Software The HBOP program requires a protein file with a Tripos Mol2 file format, and a radius file with a text file format. For the protein, hydrogen atoms need to be

Table 1. Test Set

Protein	PDB ID
Alcohol dehydrogenase	1ADF
L-Aarabinose binding protein	1ABE, 1ABF, 5ABP, 6ABP,[a] 7ABP,[a] 8ABP,[a] 1APB,[b] 1BAP,[b] 9ABP[b]
Carboxypeptidase A	1CBX, 6CPA, 7CPA, 8CPA
Cathepsin D	1LYB
Cytochrome P450cam	2CPP, 5CPP
Dihydrodipicolinate reductase	1DIH
Dihydrofolate reductase	4DFR
Endthiapepsin	1EED, 1EPO, 1EPP, 2ER6, 2ER7, 2ER9, 3ER3, 4ER1, 4ER4, 5ER2
Enolase	1EBG
D-Galactose/D-glucose binding protein	2GBP
Glucoamylase-471	1DOG
Glutamine synthetase	1LGR
Histidine binding protein	1HSL
HIV-1 protease	1AAQ, 1HBV, 1HPV, 1HTF, 1HTG, 1HVI, 5HVP, 7HVP, 9HVP
p-Hydroxybenzoate hydroxylase	2PHH
Intestinal fatty-acid binding protein	2IFB
Myoglobin	1MBI
Neuraminidase	1NNB,[c] 1NSC,[d] 1NSD[d]
Penicillopepsin	1APT, 1APU, 1PPK
Purine nucleoside phosphorylase	1ULB
Retinol binding protein	1RBP
Thermolysin	1TLP, 1TMN, 2TMN, 3TMN, 4TLN, 4TMN, 5TLN, 5TMN, 6TMN
Thrombin	1ETR, 1ETS, 1ETT
Triose phosphate isomerase	2YPI
Trypsin	1PPC, 1PPH, 3PTB, 1BRA[e]
Xylose isomerase	2XIS

a) M108L mutants. *b*) P254G mutants. *c*) Influenza virus A. *d*) Influenza virus B. *e*) D189G/G226D mutant.

added, and all other molecules (such as ligands, cofactors, water, and ions) need to be deleted. If other molecules (without the three-letter codes of the usual amino-acid residues) are included in a protein file, then the carbon atoms of the molecule are not used for the calculation. A radius file defines the radius for each Tripos atom type. The HBOP outputs two files: a grid file and a potential file. The grid file contains the coordinates of the grid points in order of their hydrophobic potential, and is written in a Tripos Mol2 file format. The grid points are grouped into 20 hydrophobic levels (ranging from HB1 to HB20), and are denoted by SYBYL colors (HB1 to HB6 are colored red, orange, yellow, green, cyan, and violet, respectively). The potential file contains the hydrophobic potential value at each grid point presented in the same order as the grid file, and is in a text file format. The HBOP is written in Fortran 90/95. The supported operating systems (OSs) are SGI IRIX, Red Hat Linux, and Apple Mac OS X. Molecular graphic software capable of reading the Tripos Mol2 file format (such as SYBYL or VMD[48,49]) is required for visualization.

HBSITE is a utility program for grid clustering, which reads the grid file as the output of the HBOP. HBSITE uses three parameters: the lowest hydrophobic level (low_{HB}), the radius of grid clustering (r_{grid}), and the minimum volume (min_{vol}; that is, the number of grid points). The default values are HB6, 1.1 Å (which is slightly larger than the default value of grid spacing), and 10 grid points, respectively. When the clustering radius is inputted manually, 0.1 Å is added to the value automatically. Before the clustering, the grid points with hydrophobic levels less than low_{HB} are rejected. If a distance between two grid points is within r_{grid}, the points are assigned to the same cluster. The grid points are then divided into clusters, and isolated grid points are eliminated. After the clustering, those with volumes larger than or equal to the min_{vol} remain. HBSITE then outputs a new grid file that contains only the grid points that belong to the clusters.

RESULTS AND DISCUSSION

When most of the ligand molecule in the experimental structure overlapped the cluster of grid points detected by HBOP/HBSITE, we considered our method to have successfully detected the binding site. The results are shown in Table 2. Our programs successfully detected the binding sites in all 26 proteins (72 complexes) using the only hydrophobic function. The ligand molecules were bound to the most hydrophobic regions in the most hydrophobic sites on all of proteins, with the exceptions of glutamine synthetase and the neuraminidase from the influenza virus B (that is, in 24 out of 26 proteins, and in 69 out of 72 complexes). It is agreement with the fact that the hydrophobic surface in the binding site is more exposed than that of the other site on a protein.[9,50–52] That is, the rate of occurrence of hydrophobic

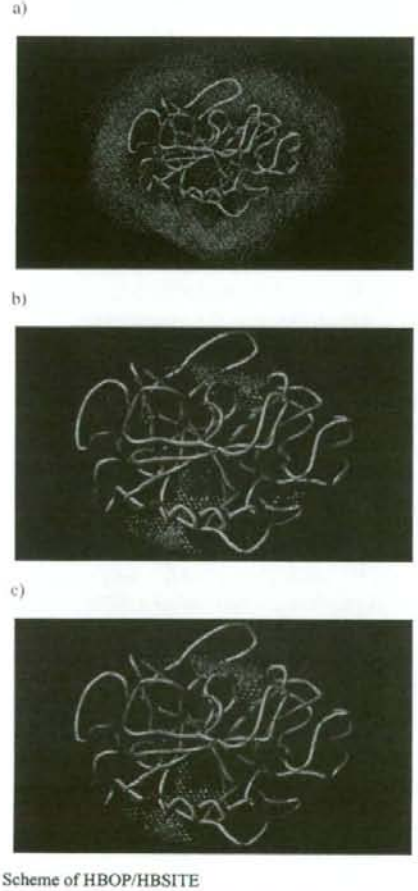


Fig. 1. Scheme of HBOP/HBSITE

(a) Grid points on a protein. (b) Hydrophobic grid points from HB1 to HB6 defined by HBOP. The grid points are represented in order of hydrophobicity as follows: red (HB1, highest), orange (HB2), yellow (HB3), green (HB4), cyan (HB5), and violet (HB6, lowest). (c) Hydrophobic sites by HBSITE.

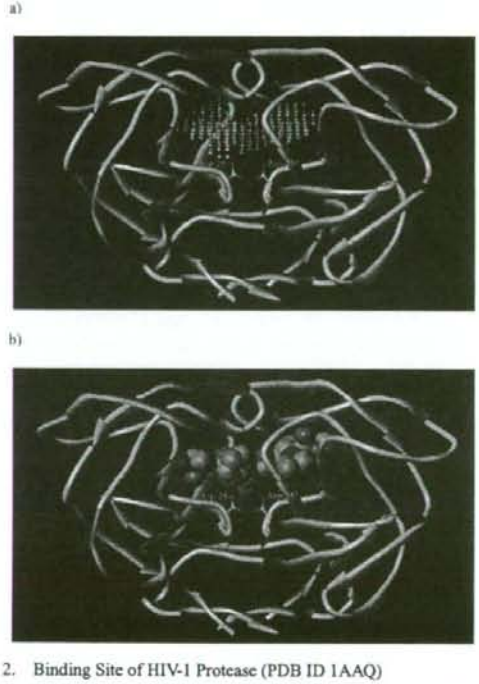


Fig. 2. Binding Site of HIV-1 Protease (PDB ID 1AAQ)

(a) One hydrophobic site detected by HBOP/HBSITE. The site is represented in order of hydrophobicity as follows: red (highest), orange, yellow, green, cyan, and violet (lowest). (b) The ligand in the binding site.

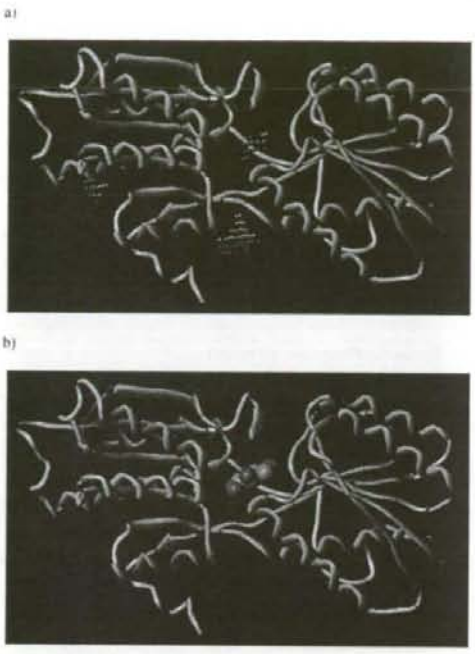


Fig. 3. Binding Site of D-Galactose/D-Glucose Binding Protein (PDB ID 2GBP)

(a) Three hydrophobic sites detected by HBOP/HBSITE. The sites are represented in order of hydrophobicity as follows: red (highest), orange, yellow, green, cyan, and violet (lowest). (b) The ligand in the binding site.

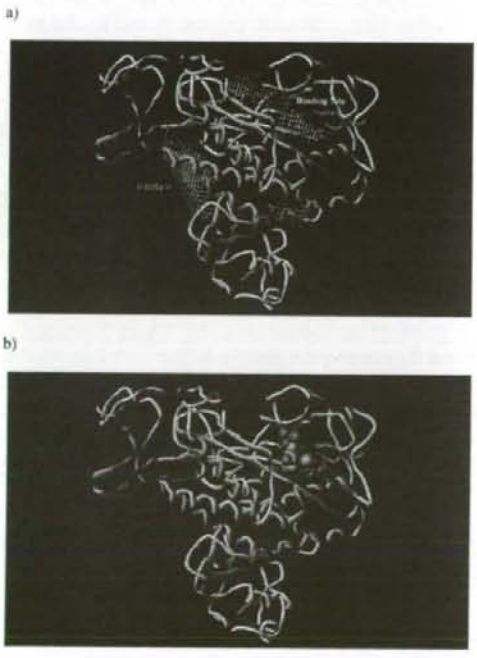


Fig. 4. Binding Site of Glutamine Synthetase (PDB ID 1LGR)

(a) Five hydrophobic sites detected by HBOP/HBSITE. The sites are represented in order of hydrophobicity as follows: red (highest), orange, yellow, green, cyan, and violet (lowest). (b) The ligand in the binding site.

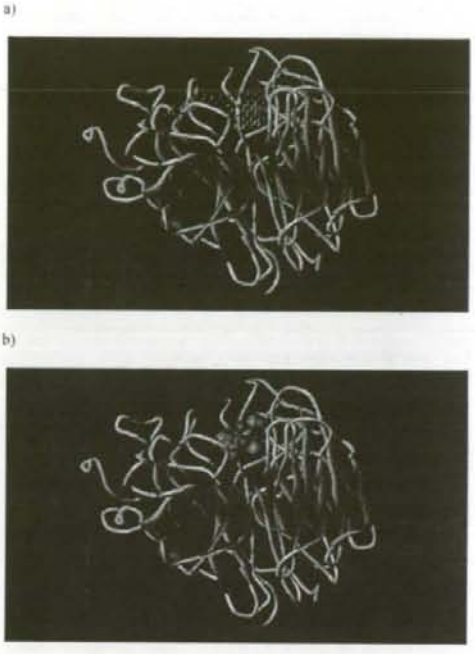


Fig. 5. Binding Site of Neuraminidase from Influenza Virus A (PDB ID 1NNB)

(a) Two hydrophobic sites detected by HBOP/HBSITE. The sites are represented in order of hydrophobicity as follows: red (highest), orange, yellow, green, cyan, and violet (lowest). (b) The ligand in the binding site.

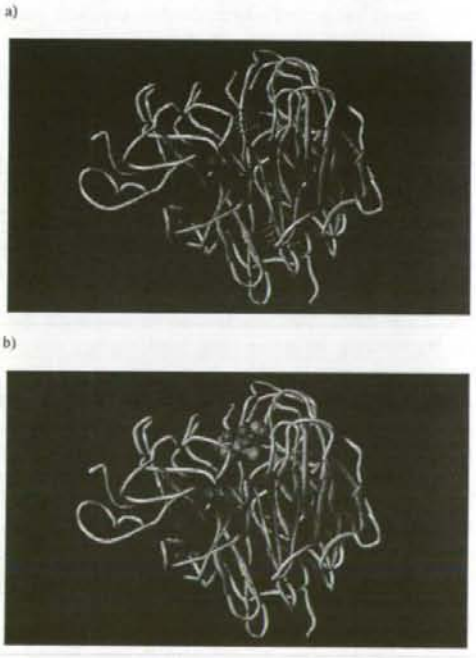


Fig. 6. Binding Site of Neuraminidase from Influenza Virus B (PDB ID 1NSC)

(a) Three hydrophobic sites detected by HBOP/HBSITE. The sites are represented in order of hydrophobicity as follows: red (highest), orange, yellow, green, cyan, and violet (lowest). (b) The ligand in the binding site.

Table 2. Detection of Ligand-Binding Sites on Proteins Using HBOP

Number of sites	Protein (Substrate and cofactor)	Detection	Protein (Substrate and cofactor)	Detection
1	Alcohol dehydrogenase (Alcohol+nucleotide)	++	Cathepsin D (Peptide)	++
	Dihydrofolate reductase (Folate+nucleotide)	++	Glucoamylase-471 (Sugar)	++
	HIV-1 protease (Peptide)	++	Intestinal fatty-acid binding protein (Fatty acid)	++
	Myoglobin (Oxygen+heme)	++	Retinol binding protein (Retinol)	++
2	Cytochrome P450cam (Camphor+heme)	++	Endthiapepsin (Peptide)	++
	Histidine binding protein (Amino acid)	++	*p*-Hydroxybenzoate hydroxylase (*p*-Hydroxybenzoic acid+nucleotide)	++
	Neuraminidase (Sugar)	+	Thermolysin (Peptide)	++
3	L-Arabinose binding protein (Sugar)	++	Enolase (2-Phospho-D-glycerate)	++
	D-Galactose/D-glucose binding protein (Sugar)	++	Penicillopepsin (Peptide)	++
	Trypsin (Peptide)	++	Xylose isomerase (Sugar)	++
4	Purine nucleoside phosphorylase (Nucleoside)	++		
5	Glutamine synthetase (Amino acid+nucleotide)	+	Thrombin (Peptide)	++
	Triose phosphate isomerase (D-Glyceraldehyde 3-phosphate)	++		
6	Carboxypeptidase A (Peptide)	++	Dihydrodipicolinate reductase (Dihydrodipicolinate+nucleotide)	++

+: ligand-binding site detected by HBOP. ++: ligand-binding site detected by HBOP and ligand bound to the most hydrophobic region.

amino acids at the binding site is relatively high.[51] When only one hydrophobic site was detected, the ligand occupied the most hydrophobic (red-orange) region in the site (Fig. 2). When a few hydrophobic sites existed on a protein, the ligand was bound to the most hydrophobic site, and occupied the most hydrophobic region on the binding site (Fig. 3). However, for glutamine synthetase (1LGR), and neuraminidases from influenza virus B (1NSC and 1NSD), the ligand molecules were bound to the secondary hydrophobic site. The biological unit of glutamine synthetase from bacteria is a dodecamer.[53] The most hydrophobic site was the interface between monomers, which was the binding site not for the small molecule but rather for the peptide chain (Fig. 4). In other words, the most hydrophobic site of glutamine synthetase was not useful for drug design. Thus, in the case of multimeric proteins, not only the most hydrophobic site but also some other sites appeared to be important for drug design. By contrast, influenza neuraminidase is a monomeric protein. For the neuraminidase from influenza virus A (Fig. 5), 1NNB, HBOP detected only upper hydrophobic sites containing the binding site. However, HBOP found new lower hydrophobic sites for the neuraminidase from influenza virus B (Fig. 6), where the secondary site was the binding site (the upper site). The similarity between the neuraminidases of influenza virus A and influenza virus B was low (sequence identity=30.8%). Numerous mutations in the neuraminidases of these influenza viruses might be responsible for the discrepancies in site detection.

HBOP was used to examine the effects of fluctuations in protein structures caused by different ligands bound to the same site and by mutations, using different structures for the same proteins (Table 3). With the exception of neuraminidase, this method was able to detect the binding site despite structural differences ranging from 0.11 Å (thermolysin, 1TLP *versus* 2TMN) to 2.42 Å (thrombin, 1ETR *versus* 1ETS) of the root-mean-square deviations (RMSDs) of the main chain. We thus considered HBOP to be relatively stable to fluctuations and mutations.

Our findings revealed that even glucose, which has many hydrogen-bonding functional groups, bound to the most hydrophobic site on the protein (Fig. 3). In Table 2, HBOP successfully detected the binding sites on all 5 sugar binding proteins. This could be explained based on the fact that glucose is more hydrophobic than water itself. Firstly, a ligand reaches a binding site by diffusion, and water molecules are excluded from the binding site.[54] The complex might be stabilized by entropy gains of the release of bound water molecules, which is the origin of the hydrophobic effect.[55] Part of the entropy gains of the release of bound water molecules could compensate for the entropy losses of the ligand and protein associated with ligand binding. In addition, the enthalpy gains of undirected van der Waals interactions between the ligand and the protein might contribute to the stabilization of the complex. Secondly, the ligand is anchored by electrostatic interactions and hydrogen bonds. The formation of hydrogen bonds and electrostatic interactions out of contact with water is an energetically favorable process.[56] Viewed from the perspective of continuum electrostatics, the environment of the binding site becomes less dielectric as it becomes more hydrophobic. Therefore, the electrostatic interactions and hydrogen bonds are strengthened.

For the majority of the proteins tested, the ligand occupied the most hydrophobic region of the binding site (Table 2). This supported the notion that the ligand interacts with the hydrophobic surface on the binding site of a protein.[9,50–52] Hydrophobic-induced acid-dissociation constant (pK_a) shifts of catalytic residues are important for enzyme activities.[57,58] In fact, the catalytic residues, Asp-25 and Asp-25′ of the

Table 3. Main Chain Fluctuations of Proteins

Protein	PDB ID								
	Reference	RMSD of main chain (Å)							
L-Arabinose binding protein	1ABE	1ABF 0.18	5ABP 0.21	6ABP[a] 0.19	7ABP[a] 0.21	8ABP[a] 0.40	1APB[b] 0.25	1BAP[b] 0.18	9ABP[b] 0.31
Carboxypeptidase A	1CBX	6CPA 0.32	7CPA 0.39	8CPA 0.41					
Cytochrome P450cam	2CPP	5CPP 0.12							
Endthiapepsin	1EED	1EPO 0.48	1EPP 0.17	2ER6 0.60	2ER7 0.20	2ER9 0.60	3ER3 0.24	4ER1 0.18	4ER4 0.60	5ER2 0.22
HIV-1 protease	1AAQ	1HBV 0.44	1HPV 0.43	1HTF 0.50	1HTG 0.68	1HVI 0.73	5HVP 0.74	7HVP 0.73	9HVP 0.52	
Neuraminidase	1NNB[c]	1NSC[d] 2.35	1NSD[d] 2.36							
Penicillopepsin	1APT	1APU 0.24	1PPK 0.12							
Thermolysin	1TLP	1TMN 0.16	2TMN 0.11	3TMN 0.11	4TLN 0.17	4TMN 0.14	5TLN 0.19	5TMN 0.12	6TMN 0.13	
Thrombin	1ETR	1ETS 2.42	1ETT 0.59							
Trypsin	1PPC	1PPH 0.17	3PTB 0.28	1BRA[e] 0.43						

a) M108L mutants. *b*) P254G mutants. *c*) Influenza virus A. *d*) Influenza virus B. *e*) D189G/G226D mutant.

HIV-1 protease were located in the most hydrophobic region of the binding site (Fig. 2). This implies that the grid points generated by HBOP can be used to define a spatial query of the hydrophobic effect for computational docking.

The assembly of a multimeric protein is the next challenge for predictive methods of the 3D structure of proteins.[59,60] It is difficult to determine whether one protein associates with another, and to identify the interfaces of the proteins. Hydrophobic interactions have also been reported to be important for protein–protein binding.[61,62] In the case of glutamine synthetase, HBOP was able to detect the interface between monomers (Fig. 4). Therefore, it might be useful as a prediction tool for protein–protein complexes.

We developed and tested a new method, HBOP, for determining the binding site of a protein using the long-range hydrophobic potential as defined by Israelachvili and Pashley. The importance of the hydrophobic interaction in ligand–protein binding was confirmed by our results. This method is a useful tool for structure-based drug design and the *in silico* screening of compounds. HBOP will be more extensively evaluated and compared with other methods in our next paper.

REFERENCES AND NOTES

1) Present address: *Laboratory of Physical Chemistry, Faculty of Pharmaceutical Sciences, Tohoku Pharmaceutical University; 4–4–1 Komatsushima, Aoba-ku, Sendai, Miyagi 981–8558, Japan.*
2) Connolly M. L., *Science*, **221**, 709—713 (1983).
3) Connolly M. L., *J. Mol. Graph.*, **11**, 139—141 (1993).
4) Ho C. M. W., Marshall G. R., *J. Comput.-Aided Mol. Des.*, **4**, 337—354 (1990).
5) Kleywegt G. J., Jones T. A., *Acta Cryst.*, **D50**, 178—185 (1994).
6) Liang J., Edelsbrunner H., Woodward C., *Protein Sci.*, **7**, 1884—1897 (1998).
7) Liang J., Edelsbrunner H., Fu P., Sudhakar P. V., Subramaniam S., *Proteins*, **33**, 1—17 (1998).
8) Brady G. P., Jr., Stouten P. F. W., *J. Comput.-Aided Mol. Des.*, **14**, 383—401 (2000).
9) Kelly M. D., Mancera R. L., *J. Med. Chem.*, **48**, 1069—1078 (2005).
10) Goodford P. J., *J. Med. Chem.*, **28**, 849—857 (1985).
11) Boobbyer D. N. A., Goodford P. J., McWhinnie P. M., Wade R. C., *J. Med. Chem.*, **32**, 1083—1094 (1989).
12) Wade R. C., Clark K. J., Goodford P. J., *J. Med. Chem.*, **36**, 140—147 (1993).
13) Wade R. C., Goodford P. J., *J. Med. Chem.*, **36**, 148—156 (1993).
14) Kuntz I. D., Blaney J. M., Oatley S. J., Langridge R., Ferrin T. E., *J. Mol. Biol.*, **161**, 269—288 (1982).
15) Shoichet B. K., Bodian D. L., Kuntz I. D., *J. Comput. Chem.*, **13**, 380—397 (1992).
16) Meng E. C., Shoichet B. K., Kuntz I. D., *J. Comput. Chem.*, **13**, 505—524 (1992).
17) Laurie A. T. R., Jackson R. M., *Bioinformatics*, **21**, 1908—1916 (2005).
18) Halgren T., *Chem. Biol. Drug Des.*, **69**, 146—148 (2007).
19) Miranker A., Karplus M., *Proteins*, **11**, 29—34 (1991).
20) Caflisch A., Miranker A., Karplus M., *J. Med. Chem.*, **36**, 2142—2167 (1993).
21) Caflisch A., Karplus M., *Perspect. Drug Discov. Des.*, **3**, 51—84 (1995).
22) Bitetti-Putzer R., Joseph-McCarthy D., Hogle J. M., Karplus M., *J. Comput.-Aided Mol. Des.*, **15**, 935—960 (2001).
23) Hansch C., *Acc. Chem. Res.*, **2**, 232—239 (1969).
24) Hansch C., Clayton J. M., *J. Pharm. Sci.*, **62**, 1—21 (1973).
25) Audry E., Dubost J.-P., Colleter J.-C., Dallet P., *Eur. J. Med. Chem.*, **21**,

71—72 (1986).
26) Audry E., Dallet Ph., Langlois M. H., Colleter J. C., Dubost J. P., *Prog. Clin. Biol. Res.*, **291**, 63—66 (1989).
27) Furet P., Sele A., Cohen N. C., *J. Mol. Graph.*, **6**, 182—189, 197—200 (1988).
28) Fauchère J.-L., Quarendon P., Kaetterer L., *J. Mol. Graph.*, **6**, 202—206 (1988).
29) Kellogg G. E., Semus S. F., Abraham D. J., *J. Comput.-Aided Mol. Des.*, **5**, 545—552 (1991).
30) Wireko F. C., Kellogg G. E., Abraham D. J., *J. Med. Chem.*, **34**, 758—767 (1991).
31) Kellogg G. E., Joshi G. S., Abraham D. J., *Med. Chem. Res.*, **1**, 444—453 (1992).
32) Meng E. C., Kuntz I. D., Abraham D. J., Kellogg G. E., *J. Comput.-Aided Mol. Des.*, **8**, 299—306 (1994).
33) Gaillard P., Carrupt P.-A., Testa B., Boudon A., *J. Comput.-Aided Mol. Des.*, **8**, 83—96 (1994).
34) Israelachvili J. N., Pashley R. M., *Nature* (London), **300**, 341—342 (1982).
35) Israelachvili J. N., Pashley R. M., *J. Colloid Interface Sci.*, **98**, 500—514 (1984).
36) Pashley R. M., McGuiggan P. M., Ninham B. W., Evans D. F., *Science*, **229**, 1088—1089 (1985).
37) Claesson P. M., Blom C. E., Herder P. C., Ninham B. W., *J. Colloid Interface Sci.*, **114**, 234—242 (1986).
38) Christenson H. K., Claesson P. M., *Science*, **239**, 390—392 (1988).
39) Christenson H. K., Claesson P. M., Berg J., Herder P. C., *J. Phys. Chem.*, **93**, 1472—1478 (1989).
40) Christenson H. K., Fang J., Ninham B. W., Parker J. L., *J. Phys. Chem.*, **94**, 8004—8006 (1990).
41) Helm C. A., Israelachvili J. N., McGuiggan P. M., *Biochemistry*, **31**, 1794—1805 (1992).
42) Meyer E. E., Lin Q., Hassenkam T., Oroudjev E., Israelachvili J., *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 6839—6842 (2005).
43) Meyer E. E., Rosenberg K. J., Israelachvili J., *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 15739—15746 (2006).
44) Eldridge M. D., Murray C. W., Auton T. R., Paolini G. V., Mee R. P., *J. Comput.-Aided Mol. Des.*, **11**, 425—445 (1997).
45) "SYBYL, 6.8.1," Tripos, Inc., St. Louis, 2002.
46) Weiner S. J., Kollman P. A., Case D. A., Singh U. C., Ghio C., Alagona G., Profeta S. Jr., Weiner P., *J. Am. Chem. Soc.*, **106**, 765—784 (1984).
47) Weiner S. J., Kollman P. A., Nguyen D. T., Case D. A., *J. Comput. Chem.*, **7**, 230—252 (1986).
48) Humphrey W., Dalke A., Schulten K., *J. Mol. Graph.*, **14**, 27—28, 33—38 (1996).
49) "VMD, 1.8.6," Theoretical and Computational Biophysics Group, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana-Champaign, 2007: ⟨http://www.ks.uiuc.edu/Research/vmd/⟩.
50) Petsko G. A., Ringe D., "Protein Structure and Function," New Science Press Ltd., London, 2004.
51) Soga S., Shirai H., Kobori M., Hirayama N., *J. Chem. Inf. Model.*, **47**, 400—406 (2007).
52) Soga S., Shirai H., Kobori M., Hirayama N., *J. Chem. Inf. Model.*, **47**, 2287—2292 (2007).
53) Yanchunas J., Jr., Dabrowski M. J., Schurke P., Atkins W. M., *Biochemistry*, **33**, 14949—14956 (1994).
54) Giovambattista N., Lopez C. F., Rossky P. J., Debenedetti P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 2274—2279 (2008).
55) Leavitt S., Freire E., *Curr. Opin. Struct. Biol.*, **11**, 560—566 (2001).
56) Chalikian T. V., *Biopolymers*, **70**, 492—496 (2003).
57) Inoue M., Yamada H., Yasukochi T., Kuroki R., Miki T., Horiuchi T., Imoto T., *Biochemistry*, **31**, 5545—5553 (1992).
58) Urry D. W., Gowda D. C., Peng S., Parker T. M., Jing N., Harris R. D., *Biopolymers*, **34**, 889—896 (1994).
59) Janin J., Henrick K., Moult J., Eyck L. T., Sternberg M. J. E., Vajda S., Vakser I., Wodak S. J., *Proteins*, **52**, 2—9 (2003).
60) Archakov A. I., Govorun V. M., Dubanov A. V., Ivanov Y. D., Veselovsky A. V., Lewi P., Janssen P., *Proteomics*, **3**, 380—391 (2003).
61) Komatsu K., Kurihara Y., Iwadate M., Takeda-Shitaka M., Umeyama H., *Proteins*, **52**, 15—18 (2003).
62) Berchanski A., Shapira B., Eisenstein M., *Proteins*, **56**, 130—142 (2004).