J. Biochem. 128, 561–574 (2000)

EmPLiCS: An Empirical Approach for Structure-Based Design of Natural Peptide Drugs[1]

Hirokazu Ishida, Tsuyoshi Shirai,[2] Yoshiko Matsuda, Yuji Kato, Masanori Ohno, Tomoya Isaji, and Takashi Yamane

Department of Biotechnology and Biomaterial Chemistry, Graduate School of Engineering, Nagoya University, Chikusa-ku, Nagoya 464-8603

Received April 7, 2000; accepted July 17, 2000

The computer implementation of a peptide drug-design strategy has been developed. The system is named EmPLiCS (Empirical Peptide Ligand Construction System) according to the strategy of the system, which searches for peptide-ligand structures by referring to empirical rules that are derived from known protein 3D structures. The system was tested on several known peptide-protein complexes. The results demonstrated the ability of this system to detect key residues of peptides that are crucial for interaction with their specific proteins. The system also showed the ability to detect the main chain trace of these peptides. Some of the main chain atoms were detected even though the complete primary structures were not reproduced, suggesting that main chain structure is important in peptide–protein recognition. The results of the present study demonstrated that the empirical rules-based system can generate significant information for use in the design of natural peptide drugs.

Key words: computer program, peptide drug, protein evolution, rational drug design.

Many medical drugs exert their pharmacological effects through tight binding to proteins or nucleic acids (*1–4*). The most important attribute of drug molecules is their recognition of particular structures of biological macromolecules. Structure-based drug design is the method of finding and improving molecular structures that complement the functional sites of target macromolecules based on their three-dimensional structures (*5–7*).

Molecules of various chemical natures have been used as the lead molecules in the reported structure-based design, including sugars (*8*), nucleic acids (*9*, *10*), and peptides (*11*). Among the variety of drug molecules, natural peptides, which are composed of the twenty biological amino acids, have considerable advantages compared to others. First, there is little difficulty in synthesizing designed molecules because the necessary techniques have been developed (*12*). Second, peptide drugs are less toxic. Third, peptide drugs can be encoded into genes (*13*), which implies that peptide drugs may be produced by cellular machineries. The tempo and place of production can be regulated by the cellular machineries for gene expression and product localization, which will work in a built-in drug-delivery system. This is the most prominent character of natural peptide drugs, because no other type of molecule can be used in this way. The structure-based design strategy, which is specifically aimed at natural peptide drugs, has significant potential applications.

Based on the drastically increased number of known protein 3D structures (*14*), the targeting of natural peptides allows the use of a unique approach, namely, purely knowledge-based methods. One use of the database is to analyze statistically the geometry of amino acid residues to evaluate the stability of peptide–protein complexes. Computer implementations of this idea have been reported.

For example, the program GEMINI accumulates information on the spatial arrangement of residues from the database, and extrapolates it in order to predict the position of ligand residues (*15*). Later, this idea was improved upon and implemented in the program X-SITE (*16*). This program collects information on the spatial arrangement at atomic resolution and renders this information into a quantitative scoring system in order to evaluate the stability of interactions. The same method was used for DNA–protein complexes using known DNA–protein complex structures from the database (*17*). The program GROW is also specialized for peptide drug design (*18*). This program uses database-derived rotamers of amino acids to generate structures of peptides. The programs LUDI and PRO_LIGAND can be also applied for natural peptide design, if the fragment libraries are restricted to that of natural amino acids (*19–21*). In a recent study, design of peptide ligand for human auto-antibody was performed using a neural network model, which represents an another empirical approach to drug design. The peptide sequences known to bind to the target protein were used as a learning set in order to deduce new sequences (*22*). Although these knowledge-based approaches to drug design and ligand prediction have been shown to work effectively, the empirical rules have not been organized into a system that can perform every required step of *de novo* design.

[1] This work was partly supported by Grants-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" (12208006) and Encouragement of Young Scientists (12780491) from the Ministry of Education, Science, Sports and Culture of Japan.

[2] To whom correspondence should be addressed. Phone: +81-52-789-3341, FAX: +81-52-789-3218, E-mail: i45282a@nucc.cc.nagoya-u.ac.jp

Current rapid expansion of sequence and structure databases of biological macromolecules is prompting the application of empirical rules for prediction of structure and function of the molecules (*23–28*). By using natural solutions derived from the databases, empirical methods can avoid massive computations which are necessary for the methods totally based on the physicochemical principles. Protein ligand prediction is also an important target of the empirical approaches. However, the performance of empirical methods in peptide ligand prediction has not yet been fully described. To what extent can a design strategy that is based totally on empirical rules detect peptide ligand structures? We have constructed a computer program that performs peptide-drug design by exclusively referring to the statistics from the protein databank. The newly developed system is called EmPLiCS (Empirical Peptide Ligand Construction System). The algorithm and performance of EmPLiCS on several known peptide–protein complexes are reported herein.

MATERIALS AND METHODS

Overview of the EmPLiCS System—A schematic overview of EmPLiCS is presented in Fig. 1. The system consists of programs for construction of empirical rules and programs for peptide design. The empirical rule construction process requires a set of known protein 3D structures as input, and generates three sets of rules: empirical peptide-peptide potential (EP^3), empirical rotamer representatives (ER^2), and empirical fitness function (EF^2). The process of peptide design requires a target protein 3D structure as an input. First, the system performs docking simulation of amino acids to a target site (seed-finding). The amino acids positioned on the protein (seeds) are used as scaffolds, upon which other residues are added during the subsequent design process (peptide-breeding). The programs were written in the C language and installed on an R8000 ONIX workstation (Silicon Graphics).

Empirical Peptide-Peptide Potential—Empirical potential energy fields were derived from 86 non-redundant protein 3D structures that were determined to a resolution higher than 2.0 Å with a crystallographic *R*-factor better than 20% using X-ray crystallography. The PDB entries are as follows (the chain IDs are presented in parentheses): 1amp, 1arb, 1ast, 1ayh, 1btc, 1cdg, 1cmb(A), 1cox, 1cpc(A), 1cpc(B), 1cse(E), 1csh, 1dbs, 1ddt, 1dri, 1ede, 1fkf, 1gcg, 1gd1(O), 1gdi, 1gof, 1gpr, 1hfc, 1hml, 1hne(E), 1hoe, 1hsl(A), 1hvl(A), 1lts(A), 1lts(D), 1mol(A), 1nar, 1nhp, 1npk, 1nsc(A), 1olb(A), 1omd, 1oya, 1paz, 1pbe, 1pcy, 1poc, 1rms, 1sar(A), 1snc, 1tca, 1tfg, 1ubq, 1xyz(A), 1ycc, 256b(A), 2acu, 2alp, 2ca2, 2cdv, 2cmd, 2cpl, 2cut, 2er7(E), 2fb4(H), 2fcr, 2gst(A), 2lal(A), 2mnr, 2pia, 2rn2, 2sic(I), 2tsc(A), 3dni, 3grs, 3il8, 3mds(A), 4enl, 4fd1, 4fxn, 4pep, 4pti, 4tnc, 5p21, 5rub(A), 5tim(A), 6tmn(E), 7aat(A), 7xia, 8dfr, and 8rsa(A).

The empirical potential field was derived from statistics of the spatial distribution of atoms in the protein molecules. This method is similar to that used in the X-SITE program (*16*) except for the following differences. In this method, protein structures are divided into rigid atom groups, which are called proto-groups. The proto-groups are defined so that each defines a unique coordinate system. An atom that has two or more covalent bonds with non-hydrogen atoms defines a 3D coordinate system. Each proto-group contains at least one such atom and, as long as the atoms are not connected with a single covalent bond, a proto-group contains as many atoms as possible (Table I). If the

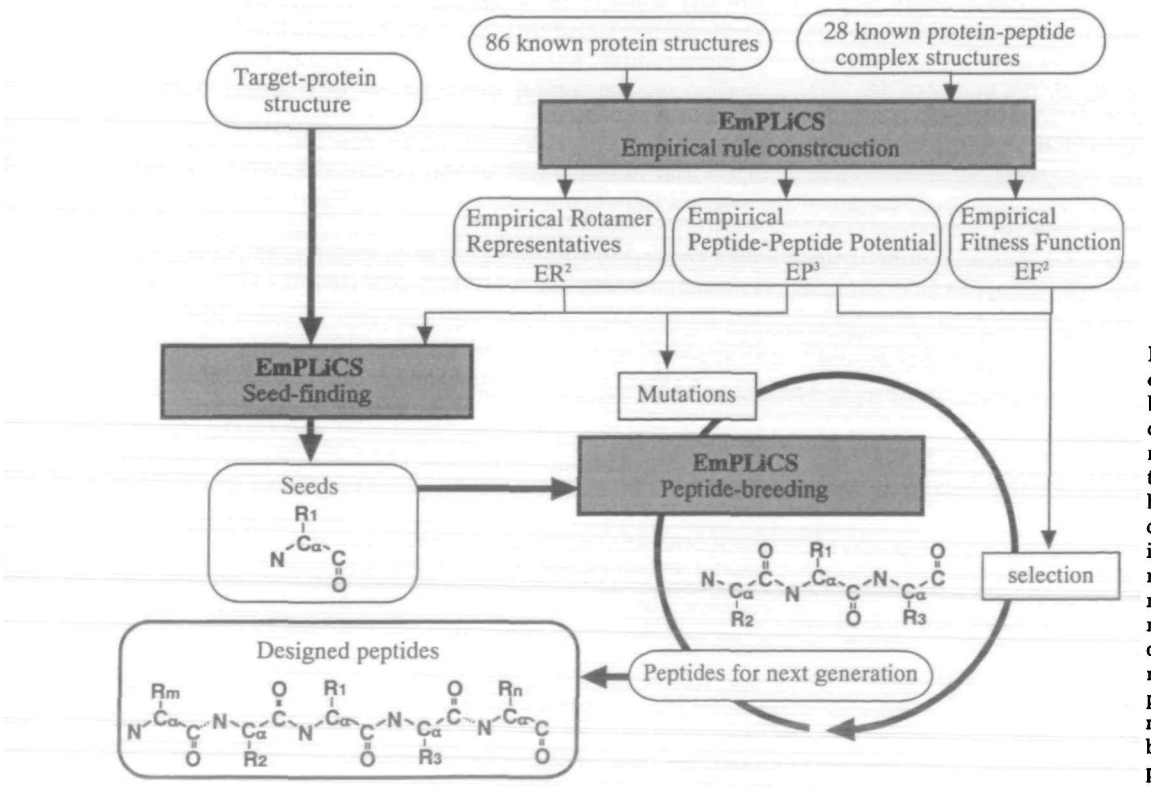

Fig. 1. **A schematic overview of EmPLiCS.** The meshed boxes indicate the three processes executed by the system, namely, empirical rule construction, seed-finding, and peptide-breeding. The open boxes are operations in the peptide-breeding process. The three empirical rules are represented by meshed round-boxes. The open round-boxes indicate the input or output data (atom coordinates of amino acids, peptide, or proteins). Examples of peptide model in seed-finding, peptide-breeding and final (designed peptide) stages are presented.

TABLE I. **EP[3] and ER[1] statistics.**

Residue	No. samples	No. rotamers	Proto-group definitions
Ala	1,866	1	[(C_α-C_β)-C]
Arg	838	63	[N-(C_α)-C], [C_α-(C_β)-C_γ], [C_β-(C_γ)-C_δ], [C_γ-(C_δ)-N_ε], (N_ε-[C_ζ-$N_{\eta 1}$-$N_{\eta 2}$])
Asn	1,004	6	[N-(C_α)-C], [C_α-(C_β)-C_γ], [(C_γ-$O_{\delta 1}$-$N_{\delta 2}$)]
Asp	1,291	3	[N-(C_α)-C], [C_α-(C_β)-C_γ], [(C_γ-$O_{\delta 1}$-$O_{\delta 2}$)]
Cys	286	3	[N-(C_α)-C], [C_α-(C_β-S_γ)]
Gln	755	28	[N-(C_α)-C], [C_α-(C_β)-C_γ], [C_β-(C_γ)-C_δ], [(C_δ-$O_{\varepsilon 1}$-$N_{\varepsilon 2}$)]
Glu	1,154	7	[N-(C_α)-C], [C_α-(C_β)-C_γ], [C_β-(C_γ)-C_δ], [(C_δ-$O_{\varepsilon 1}$-$O_{\varepsilon 2}$)]
Gly	1,783	1	[N-(C_α)-C]
His	446	6	[N-(C_α)-C], [C_α-(C_β)-C_γ], (C_γ-$C_{\delta 2}$-[$N_{\delta 1}$-$C_{\varepsilon 1}$-$N_{\varepsilon 2}$])
Ile	1,128	6	[N-(C_α)-C], [C_α-(C_β-$C_{\gamma 2}$)], [C_β-($C_{\gamma 1}$-$C_{\delta 1}$)]
Leu	1,616	4	[N-(C_α)-C], [C_α-(C_β)-C_γ], [(C_γ-$C_{\delta 1}$-$C_{\delta 2}$)]
Lys	1,195	67	[N-(C_α)-C], [C_α-(C_β)-C_γ], [C_β-(C_γ)-C_δ], [C_γ-(C_δ)-C_ε], [C_δ-(C_ε-N_ζ)]
Met	410	20	[N-(C_α)-C], [C_α-(C_β)-C_γ], [C_β-(C_γ)-C_δ], [C_γ-(S_δ-C_ε)]
Phe	820	6	[N-(C_α)-C], [C_α-(C_β)-C_γ], (C_γ-$C_{\delta 1}$-$C_{\delta 2}$-[$C_{\varepsilon 1}$-$C_{\varepsilon 2}$-C_ζ])
Ser	1,426	3	[N-(C_α)-C], [C_α-(C_β-O_γ)]
Thr	1,293	3	[N-(C_α)-C], [(C_β-$O_{\gamma 1}$-$C_{\gamma 2}$)]
Trp	314	6	[N-(C_α)-C], [C_α-(C_β)-C_γ], (C_γ-$C_{\delta 1}$-$C_{\delta 2}$-[$N_{\varepsilon 1}$-$C_{\varepsilon 2}$-$C_{\zeta 2}$]-$C_{\zeta 3}$-$C_{\eta 3}$-$C_{\eta 2}$)
Tyr	801	6	[N-(C_α)-C], [C^α-(C_β)-C_γ], (C_γ-$C_{\delta 1}$-$C_{\delta 2}$-[$C_{\varepsilon 1}$-$C_{\varepsilon 2}$-C_ζ]-O_η)
Val	1,491	3	[N-(C_α)-C], [(C_β-$C_{\gamma 1}$-$C_{\gamma 2}$)]
cis-Pro	60	1	(N-C_α-[C_β-C_γ-C_δ])
trans-Pro	917	1	(N-C_α-[C_β-C_γ-C_δ])
Peptide plane	20,808	225	[(N-C-O)]
N-terminus	81	1	[(N)-C_α-C]
C-terminus	86	1	[C_α-(C-O]-OXT)

Columns: Residue, names of residues; No. samples, total numbers of the residues observed in the sample proteins; No. rotamers, numbers of rotamers defined in ER[2] for the residues (combination of representative χ^1–χ^5 angles for side chains, or combination of ϕ–ψ angles for main chain); Proto-group definitions, atomic symbols enclosed in parentheses represent the proto-group, and those in square brackets are the three atoms that define reference coordinate system of the proto-group.

content of a proto-group is less than three atoms, atoms of other groups that are directly connected to the group are used to define the coordinate system.

The process of the potential field is schematically presented in Fig. 2. Each proto-group in the protein molecules is superimposed on its reference coordinate system (Step 1 in Fig. 2). Atoms around the proto-groups are categorized into sixteen types and are called target-atoms (Table II). The count of the target-atoms is accumulated on a grid system, which consists of evenly distributed points at 0.5 Å intervals in a 12 × 12 × 12 Å^3 box centered at the origin of the proto-group (Step 2 in Fig. 2).

The observed target-atom frequencies are redistributed in the grid system (Step 3 in Fig. 2). Suppose $f_{ij}(\boldsymbol{r})$ is an observed frequency of the target-atom i for the proto-group j at the grid point $\boldsymbol{r}$, and n_{ij} is the total observed number of the target-atom for the proto-group. The frequency is modified as

$$f_{ij}(\boldsymbol{r}) = \sum_{\boldsymbol{s}} f_{ij}(\boldsymbol{s}) \exp(-|\boldsymbol{r}-\boldsymbol{s}|^2/2\sigma^2)$$

where $\sigma = 0.5 + 50/n_{ij}$. When the total observed number (n_{ij}) is small, the large σ value creates a featureless distribution reflecting the low reliability of the statistics.

The modified target-atom frequencies at each grid point are converted into potential energy (Step 4 in Fig. 2). Assuming the Boltzmann distribution of the atoms around proto-groups, the potential energy (E) at a position $\boldsymbol{r}$ is defined as

$$E = -RT \ln \{f_{ij}(\boldsymbol{r})/\langle f_{ij}(\boldsymbol{r})\rangle\}$$

where $\langle f_{ij}(\boldsymbol{r})\rangle$ is the average of the frequency over the grid

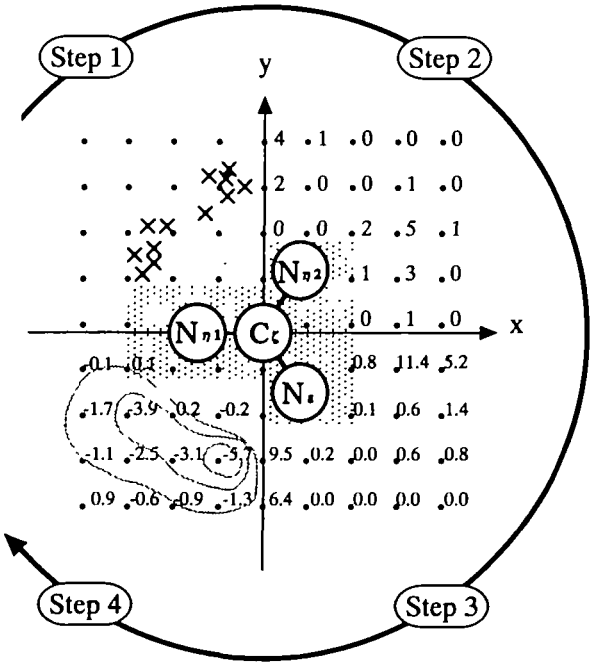

Fig. 2. **A scheme of the process of EP[3] from protein coordinates.** Step 1: Target-atoms are transferred into the reference coordinate system of the proto-group. Step 2: The counts of target-atoms are heaped up on the grid system. Step 3: The observed numbers of target-atoms are redistributed and normalized. Step 4: The frequency is converted into EP[3].

TABLE II. **Definition and sample number of target-atoms.**

Name	Formula	No. samples
Main chain		
Aliphatic-C	>CH–	20,894
Carbonyl-C	>C=	20,894
Carbonyl-O	=O	20,808
Amide-N	–NH–	20,813
Side chain		
Aliphatic-C	$-CH_3$ $-CH_2-$ >CH–	42,400
Carbonyl-C	>C=	4,204
Aromatic-C	=CH–	13,576
Carbonyl-O	=O	1,759
Carboxyl-O[a]	$-O^-$	5,044
Hydroxyl-O	–OH	3,520
Amino-N	$-NH_2$	1,759
Charged-N[b]	$-NH_3^+$	2,952
Aromatic-N[c]	–NH– =N–	2,044
Thiol-S	–SH	286
Sulfur-S	–S–	410
Water-O	HOH	19,017

Columns: Name, name of target-atom; Formula, bonding patterns of target-atom; No. samples, total number of target-atoms found in the sample proteins. [a]Carboxyl-O is the oxygen atom of Asp and Glu side chains and C-terminal carboxylate group. [b]Charged-N is the nitrogen atom of Arg (N_η), Lys and N-terminal amino group. [c]Aromatic-N is the nitrogen atom of Arg (N_ϵ), His and Trp side chains.

points. The function of potential energy is called EP^3 (empirical peptide-peptide potential).

Two different forms of EP^3 were developed, namely, rEP^3 and nEP^3, which correspond to the potentials for remotely connected and neighboring residues in primary structure, respectively. In the process of rEP^3, only the atoms that are separated by more than six residues from the residue to which the proto-group in consideration belongs are counted. The nEP^3 is derived from the atoms that are within six residues from the proto-group's residue. The rEP^3 is used for energy calculation of inter-molecular interaction, and the nEP^3 is used for intramolecular interaction.

The energy of a peptide-protein complex is calculated as the sum of the rEP^3 and nEP^3 values for the atoms in the fields defined by the peptide and protein structures. EP^3 fields of proto-groups are superimposed onto proto-groups of the protein or peptide molecules. The values from every proto-group are heaped up on a grid system that is fixed on the complex structure. The fields are synthesized for the sixteen target-atoms separately. The total energy of a peptide-protein complex structure (E_{com}) is defined as

$$E_{com} = (E^R_{pro} + E^R_{pep} + E^N_{pep})/2$$

where E^R_{pep} is the summation of the potentials for peptide atoms in the rEP^3 fields defined by the protein, and E^R_{pro} is the summation of the values for protein atoms in the rEP^3 fields defined by the peptide. The E^N_{pep} is the summation of the values for peptide atoms in the nEP^3 fields defined by the peptide itself.

Empirical Rotamer Representatives—Preference in dihedral angle was also deduced from the 86 protein structures. A total of 255 favored combinations of ϕ and φ angles were selected for main chain torsion angles. Side chain angles (χ) are treated separately, and up to three representative value(s) are selected for each angle. Rotamers of amino acid residues in the design process are prepared by combining the representative values (Table I). This dihedral angle preference is referred to as ER^2 (empirical rotamer representatives).

Empirical Fitness Function—The process of finding appropriate peptide drugs requires comparisons of stability among peptide models with a variety of sequences and lengths. In general, a peptide with more internal degrees of freedom requires more reduction of enthalpy in the binding process, because the reduction in conformational entropy is larger than that of molecules of lower complexity. An empirical fitness function that implicitly includes the conformational entropy is used in this system to normalize the difference among peptides. In the empirical fitness function, the internal degree of freedom of a peptide was simply assumed to be proportional to the number of single bonds (m) of the molecule, and the energy of peptide-protein complex was normalized as

$$E_{nor} = E_{com} - a\,m - b.$$

The constants (a and b) of the function were derived from 28 known peptide-protein complex structures (Table III). The complex structures are those determined to a resolution higher than 3.0 Å using X-ray crystallography or those determined by NMR. The EP^3 energy (E_{com}) of the 28 known peptide–protein complexes was calculated, and the constants (a and b) were obtained by a first-order regression. The result is shown in Fig. 3. The values for a and b are determined to be –38.0 and –87.3 (kJ/mol), respectively. The correlation coefficient between the number of single covalent bonds and energy was 0.93.

Seed-Finding and Peptide-Breeding Processes—Initial structures for peptide design process are single amino acids which are called seeds. Every possible combination of translation, rotation, and ER^2 of amino acid is examined within a space around a defined target site of protein molecule in order to find the seeds. This process is called seed-finding. In the subsequent process, the seeds are elongated into peptides by adding amino acid residues. This process is called peptide-breeding. In both processes, terminal groups in seeds and growing peptides are modeled as amido-N or carbonyl-C=O, since the unit of amino acid residue in this system is N-C_α(-R)-C=O (R = side chain) (Fig. 1). Side chains are composed of the proto-groups that are defined in Table I. Hydrogen atoms are not included.

Evolutionary strategy is employed to search for peptide structures; the peptide-breeding process repeats the following three steps (Fig. 1).

1) From n peptides in a generation, $2n$ peptides are prepared for the next generation. They are the ensemble of the peptides of the current generation and their mutant froms in equal numbers. The method of mutagenesis for each peptide is randomly selected from the following six operators: extension/deletion of one residue to/from N- or C-terminals, side chain replacement of a randomly selected residue by a randomly selected amino acid, and conformational modification (an alteration of a randomly selected dihedral angle). The rotamer of a residue to be added or replaced is selected by exhaustively testing ER^2 in the EP^3 fields.

2) Fitness values are calculated for the $2n$ peptides. The fitness value of a peptide is defined as

$F(i) = nE_{nor}(i)/\langle E_{nor}(i)\rangle$.

The $F(i)$ value is the number of the peptides i in the next generation.

3) The top n peptides out of $2n$ are selected for the next generation according to their $F(i)$ values, and the remaining n are excluded. The nEP^3/rEP^3 ratio of each peptide is also monitored in this step. When a peptide forms several intramolecular interactions, it can increase the $F(i)$ value without interacting with the target protein. Such a self-sufficient peptide is detected by nEP^3/rEP^3 ratio and excluded.

Applying EmPLiCS to Known Complexes—The newly developed system was tested on the seven following peptide–protein complex structures (in parenthesis are abbreviations of the protein name and PDB codes): *Salmonella typhimurium* protein methyltransferase CheR (CheR, 1bc5) (*29*), PDZ-3 domain from rat synaptic protein PSD-95 (PDZ domain, 1be9) (*30*), human class I (MHC I, 1hhg) (*31*) and class II major histocompatibility complexes (MHC II, 1dlh) (*32*), Src homology 3 domain from *Caenorhabiditis elegans* protein SEM-5 (SH3 domain, 1sem) (*33*), human immunodeficiency virus 1 protease (HIV1 protease, 7hvp) (*34*), and yeast nuclear-import factor karyopherin α (Kap α, 1bk6) (*35*).

The crystal structures were used for targets of the system in order to determine if the system can detect the structures of known peptide ligands. Two types of seeds were used in this test. One group was composed of the residues of the known peptides from the complex structures and are called "known seeds." The other group was composed of seeds that were newly sought by seed-finding and are called "sought seeds." The seed-finding was executed in a 7.0 × 7.0 × 7.0 $Å^3$ box around the C_α atoms of known peptides. Every combination of translations, rotations and rotamers (from ER^2 data base) of the amino acid was examined in the rEP^3 field synthesized on the protein structure. The step of translation was 0.6 Å in all directions and the step of spherical polar angle was 15° in the ranges of $0° \le \phi < 180°$, $0° \le \varphi < 180°$, and $0° \le \kappa < 360°$. The top 2,000 seeds found in this process were used for clustering analysis. If the root mean square deviation (rmsd) of the corresponding atoms of two seeds was less than 0.8 Å, the two seeds were accommodated into the same cluster. This procedure was repeated until no further unification occurred.

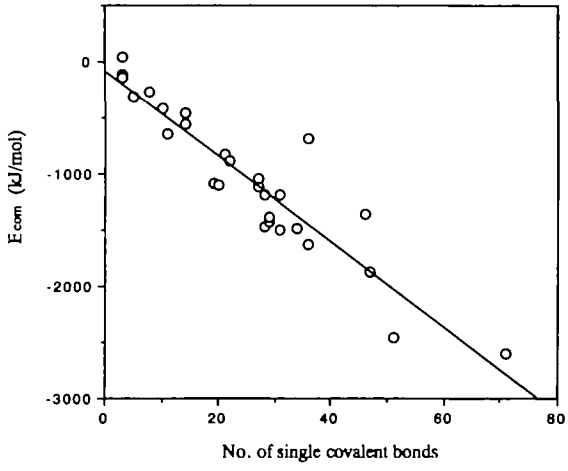

Fig. 3. **Plot of E_{com} values of the 28 known peptide-protein complexes with respect to the numbers of single covalent bonds in the peptides.** The line shows the first-order regression of the data points.

TABLE III. **Proteins for EF^3 deduction.**

Protein	Code	Peptide sequence	No. atoms	No. bonds
Allosteric chorismate mutase	1csm(L)	W	15	3
α-Thrombin	1abj	FPR-CH_2	30	10
	1fph(F)	Ace-DFLAEGGGVR-CH_2	104	47
	1fph(I)	GDFEEIPEEYLQ	75	34
cAMP-dependent protein kinase	1cmk	TTYADFIASGRTGRRNAIHD	157	71
Class I MHC	1hhg(C)	TLTSCNTSV	63	28
	1hhi(C)	GILGFVFTL	69	29
	1hhj(C)	ILKEPVHGV	70	31
	1hhk(C)	LLFGYPVYV	77	28
	1tmc	EVAPPEYHRK	86	36
	1vaa	RGYVYQGL	68	29
Class II MHC	1dlh(C)	PKYVKQNTLKLAT	106	51
Endothiapepsin	1er8	HPFHLLVY	73	27
γ-Chymotrypsin	2gct	UPGAY	29	8
His-binding protein	1hsl(D)	H	11	3
IgG1 Fab fragment	2igf	EEVVPHKK	58	27
	1ggi(P)	CKRIHIGPG	68	31
IgG2A Fab fragment	1ifh	Ace-DVPDYAS	57	20
IgG2A Fv fragment	1iif	GATPQDLNTML	80	36
Immunoglobulin λ light chain dimer	1mcb	Ace-QFHP	41	14
L-Asparaginase	3eca(E)	D	9	3
Lys, Arg, Orn-binding protein	1laf	R	12	5
Oligo-peptide binding protein	1ola	VKPG	28	11
	1olc	KKKA	33	19
SH3 domain	1cka	PPPALPPKK	65	22
	1sem(C)	Ace-PPPVPPRRR	56	14
Thioredoxin	1mdi	FRFRYVCEGPSHG	110	46
Rhizopuspepsin	3apr	PFHF-φ[CH_2-NH]-FV	56	20

Columns: protein, name of protein; Code, PDB code. In parenthesis are the chain IDs of peptides; Peptide sequence, amino acid sequences of the peptide in one letter code. Non-standard atom groups are indicated as follows: CH_2 (methylene), Ace (acetyl), and U (unknown side chain); No. atoms, number of non-hydrogen atoms in the peptides; No. bonds, number of single covalent bonds of the peptide.

Each cluster was represented by the seed that had the best EP^3 value in the cluster. The top 15 cluster-representatives were submitted for the subsequent peptide-breeding. The seed-finding was executed for twenty amino acids, and a total of 300 seeds were prepared for one protein.

The size of population was fixed to at 300 during the peptide-breeding. The nEP^3/rEP^3 ratio was limited to 1.3; peptides that violated this limit were excluded immediately. One peptide-breeding run consisted of 100 generations, and a total of five runs were executed for each protein. Accordingly, a total of 1,500 final peptides were obtained for each target-protein.

Comparison of EmPLiCS-Designed Peptides with Phage-Display Library Peptides—Peptide sequences designed for MHC II with EmPLiCS system were compared with the reported peptide sequences that were screened from phage-display libraries with the same MHC II molecule (*36*). Sequences of nonapeptide, that correspond to the fragments from N- to C-terminal anchor sites, were extracted from the EmPLiCS-designed peptides (top half of the final population of sought-seed design) by avoiding redundancy. The sequences were aligned with each of the 60 sequences from the phage-display library without gaps, and the maximum matches of residues were counted. For a statistical evaluation, the designed-sequences were randomized by conserving the amino acid compositions, and the same sequence comparison was repeated 1,000 times to obtain the values of expected distribution and standard deviation.

Peptide Modeling with Insight II/Discover Program—To evaluate the performance of EmPLiCS, the known peptides were modeled by using the commercially available software, Insight II/Discover release 98.0 (MSI/Ryoka Systems) installed on an O_2 workstation (Silicon Graphics), and the modeled peptides were compared with the EmPLiCS-designed peptides.

Known seeds were used for the initial structures, and the residues of known peptides were added to the seeds one by one; therefore, the sequences were restricted to that of the known peptides. The starting conformation of an added residue was arbitrarily determined by avoiding steric hindrance, then a stable conformation was sought by molecular-dynamics (MD) simulation. Modeling was continued until all the residues in the known peptide were modeled, or the modeling came to an apparent dead-end, that is, all of the possible conformations of an added residue caused serious steric hindrance. The coordinates of protein atoms and the already modeled part of the peptide were fixed during MD simulation, except for the amido or carbonyl groups to which the added residue was connected. Terminals were modeled as $-NH_2$ or $-C(=O)H$.

The MD simulations were executed at 1,000 K for 10 ps (1 ps = 10^{-12} s) *in vacuo*, and the trajectories were sampled every 1 ps. Then the sampled structures were annealed at 300 K for 0.1 ps and energy-minimized. The structure with the lowest energy was used for further extension of residues. The AMBER force field was used for the simulations (*37*). Every known peptide was modeled twice by using different seeds, and the better results were used for comparison.

RESULTS AND DISCUSSION

Examples of rEP^3—The program system EmPLiCS was designed to evaluate the stability of peptide–protein complexes based on empirical potential field derived from the protein databank. The energy for atom pairs that are distantly or proximally connected in the primary structure is treated using rEP^3 or nEP^3, respectively. The former is used for evaluation of intermolecular (peptide–protein) interactions and the latter is used for intramolecular (peptide–peptide) interactions.

An example of the rEP^3 field of guanidine proto-group is shown in Fig. 4a. In the fields for carbonyl-O and -C target-atoms, well-defined minima were observed at the positions appropriate for H-bonding to the $N_{\eta 1}$, $N_{\eta 2}$, and N_ϵ atoms of guanidine. The average distance of minima of carbonyl-O to the nitrogen atoms is 3.0 Å. The angles of acceptor-associated atom (carbonyl-C)–acceptor (carbonyl-O)–donor ($N_{\eta 1}$, $N_{\eta 2}$, and N_ϵ) range from 137° to 171°. The average distance between the minima for carbonyl-O and -C atoms is 1.1 Å. This example typically shows how a combination of rEP^3 fields can define the empirically favored geometry of interaction.

The rEP^3 field of phenyl proto-group for aromatic-C target-atom is shown in Fig. 4b. Three regions of lower potential were observed, that were above and below the ring and around edge of the ring, which seems to reflect the reported preference of aromatic rings for edge-to-face interactions (*38*). Compared to the example of the H-bond, the favored region of ring–ring interaction is not localized, reflecting the character of the interaction that is less sensitive to geometry.

Since rEP^3 was derived from natural protein structures, an intriguing feature was observed in the field of the Phe residue. The rEP^3 field of phenyl proto-group for hydroxyl-O target-atom is presented in Fig. 4c. Some potential minima were observed near the C_ζ atom. When the phenol proto-group of Tyr residue and its field for the same target-atom are superimposed on the Phe system, the minima for hydroxyl-O, which may H-bond to the hydroxyl group of Tyr, appeared to coincide to those observed for Phe under the same contour levels.

Phe and Tyr are the most frequently interchanging amino acids in protein evolution, when the substitution rate is normalized by the observed frequency of amino acids (*39*). Phe residues that have a proximal unsaturated H-bonding partner may have been Tyr residues previously. Tyr to Phe substitution might have retained the H-bond partner and the retained groups are observed in the EP^3 statistics. Except for this "historical interaction," visual inspection of rEP^3 fields showed that they were principally generated from a combination of H-bond, electrostatic, Van der Waals, and hydrophobic interactions.

Examples of nEP^3—Since a peptide has several rotamers, selecting the appropriate rotamer is the most difficult step of peptide design. The protein databank can be used for this purpose as a source of adequate peptide fragments. Although EmPLiCS does not use the peptide structures directly from the database, instead nEP^3 and ER^2 can perform this function. Since the nEP^3 is derived by taking only neighboring residues in the primary structure into account, nEP^3 is dominated by local interactions that are mainly used to determine the dihedral angles between neighboring proto-groups.

An example of the nEP^3 field of the phenyl proto-group for main chain aliphatic-C target-atom is shown in Fig. 5a.

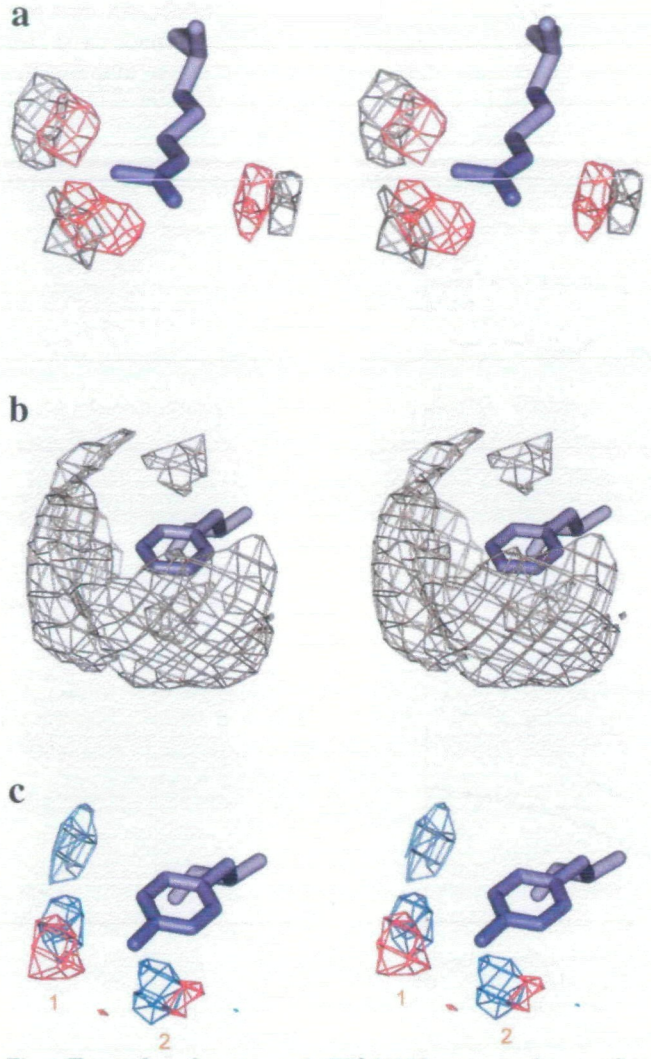

Fig. 4. **Examples of contoured rEP³ field in stereo-views.** a: Superposition of the rEP³ fields of guanidine proto-group for carbonyl-O (red) and carbonyl-C target-atoms (gray). Both of the fields are contoured at –2.0 kJ/mol level. b: The rEP³ field of phenyl proto-group for aromatic-C target-atom (gray). The field is contoured at –1.0 kJ/mol level. c: Superposition of the rEP³ fields of phenyl of Phe (red) and phenol of Tyr (blue) proto-groups for hydroxyl-O target-atom. Both of the fields are contoured at –2.2 kJ/mol level. The target-atom density at the peaks 1 and 2 (labeled in figure) for Tyr are higher than the average by 8.1 and 5.0 σ, respectively. The same density for Phe at the two peaks are 4.5 and 4.1 σ higher than the average, respectively.

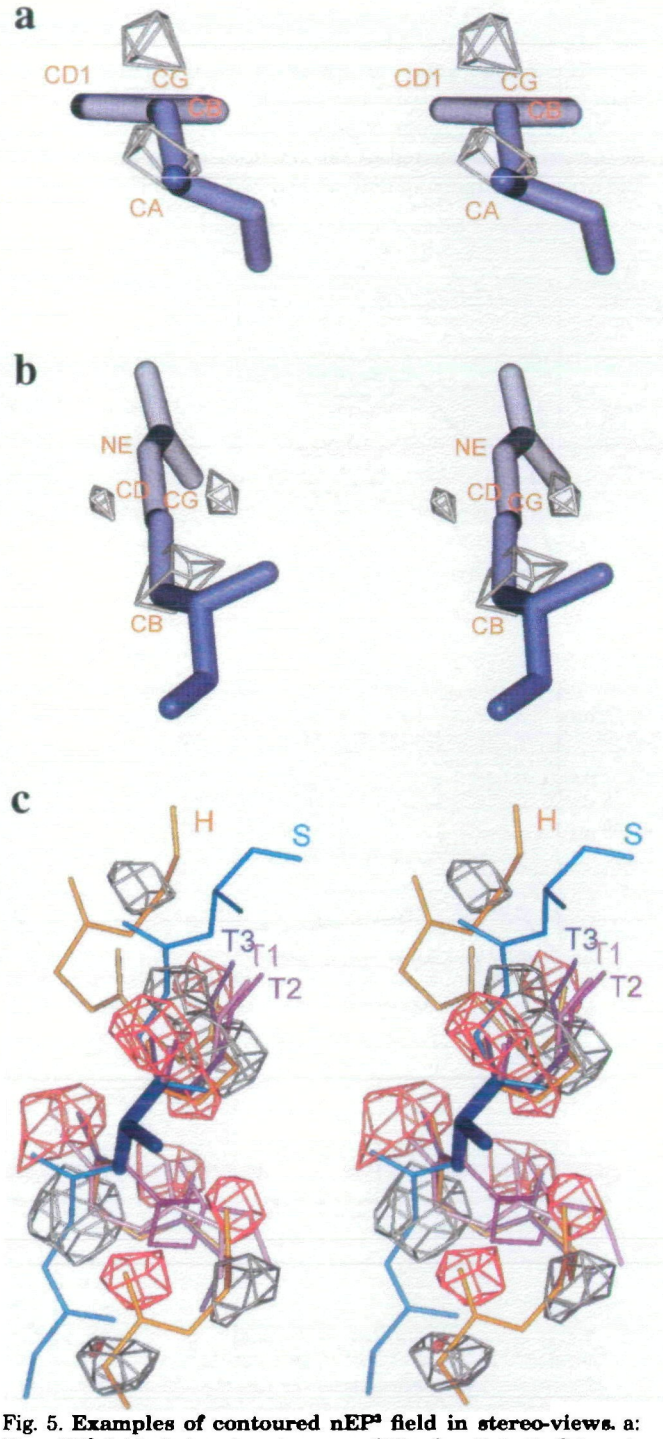

Fig. 5. **Examples of contoured nEP³ field in stereo-views.** a: The nEP³ field of phenyl proto-group of Phe for aliphatic-C target-atom. The C_β-C_γ bond is perpendicular to the surface of the paper. The field is contoured at –5.1 kJ/mol level. b: The nEP³ field of C_δ proto-group of Arg for aliphatic-C target-atom. The C_γ-C_δ bond is perpendicular to the paper. The field is contoured at –7.0 kJ/mol level. c: The nEP³ field of C_α proto-group of Ala for main chain aliphatic-C (gray) and carbonyl-O (red) target-atoms. The main chain atoms for the righthand helix (H), strand (S), type I (T1), II (T2), and III (T3) turns are superimposed on the C_α proto-group of Ala. Both of the fields are contoured at –4.0 kJ/mol level.

The two minima in the field indicate the favored position of the C_α atom, showing preferred dihedral angles $C_{\delta 1}$-C_γ-C_β-C_α (χ^2) of 90° and 270°. An example of the nEP³ field of the C_δ proto-group of Arg for aliphatic-C target-atom is shown in Fig. 5b. The three minima correspond to the angle C_β-C_γ-C_δ-N_ε (χ^3) of 0°, 120°, and 240°.

The fields of C_α proto-group (of Ala) for C_α and main-chain-O target-atoms are shown in Fig. 5c. Preference for regular secondary structures is presented by the fields. These examples show that the nEP³ field can be used for detecting favored dihedral angles.

Testing the EmPLiCS on Peptide-Protein Structures—The EmPLiCS system first searches for seed (single amino acid) positions on the target site of protein molecules using the empirical energy functions introduced above. The sys-

tem then breeds the seeds into peptides by mimicking the evolutionary process. These processes were tested on seven known peptide–protein complexes to see if the system could detect the peptide ligand structures.

Two different groups of seeds were used in this test. One of the groups consisted of the residues of peptide found in the known complex structures (known seeds) and the other consisted of seeds that were newly sought around the binding sites by seed-finding (sought seeds). In both cases, the initial generation was composed of 300 single amino acids, and resulted in the same number of peptides. The execution time for one peptide-breeding process ranged from 6 to 10 h.

Progress in the peptide-breeding process is presented in

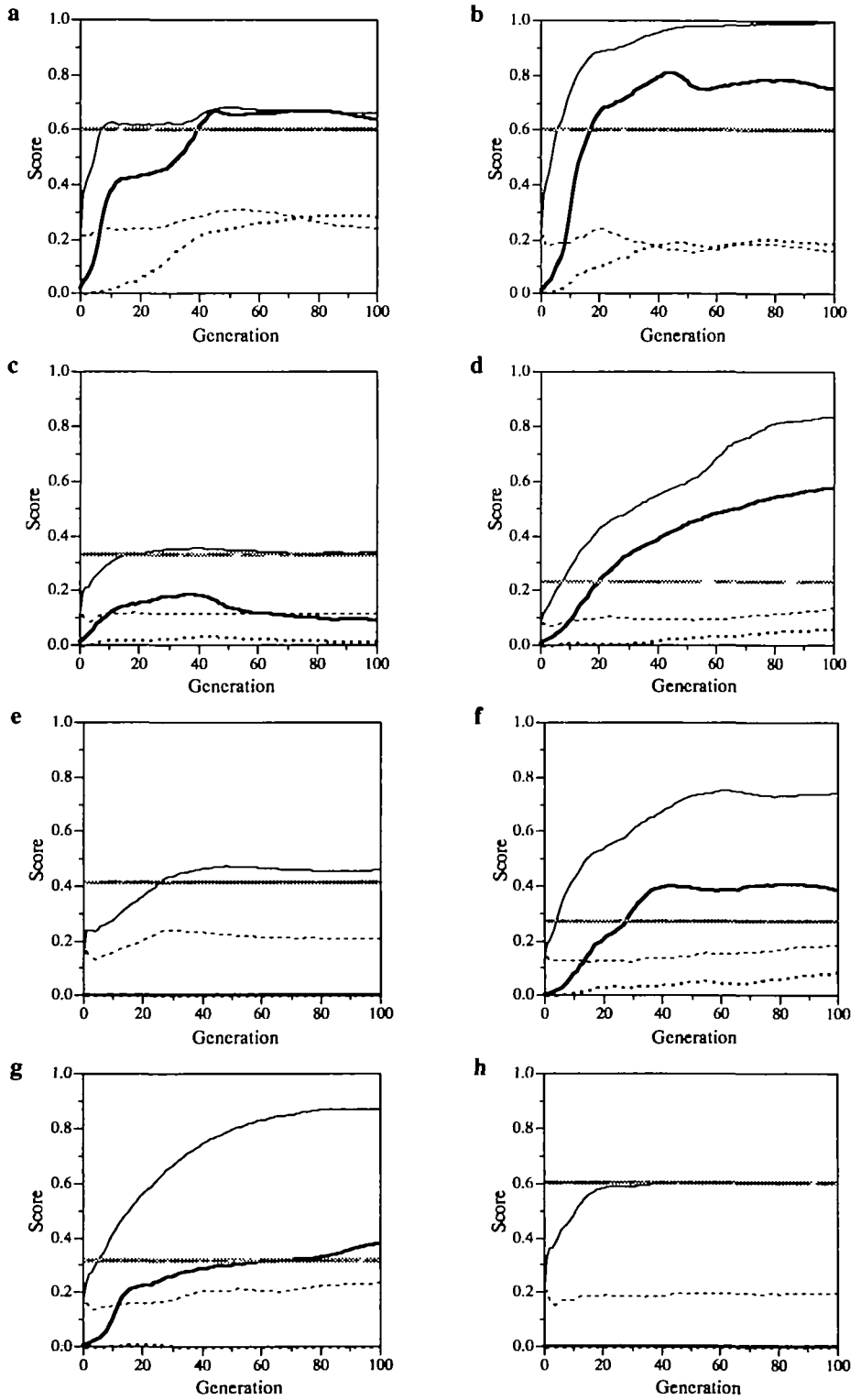

Fig. 6. **Plots of average scores of peptides during peptide-breeding with respect to generation number.** In each panel, thin and thick lines show the progress in breeding from known and sought seeds, respectively, and solid and dotted lines correspond to scores for detection of main chain trace and sequence (defined in the text), respectively. Gray-horizontal lines show the scores of main chain trace detection of the peptides modeled by using InsightII/Discover program. a: CheR. b: PDZ domain. c: MHC I. d: MHC II. e: SH3 domain. f: HIV1 protease. g: Larger NLS-binding site of Kap α. h: Smaller NLS-binding site of Kap α.

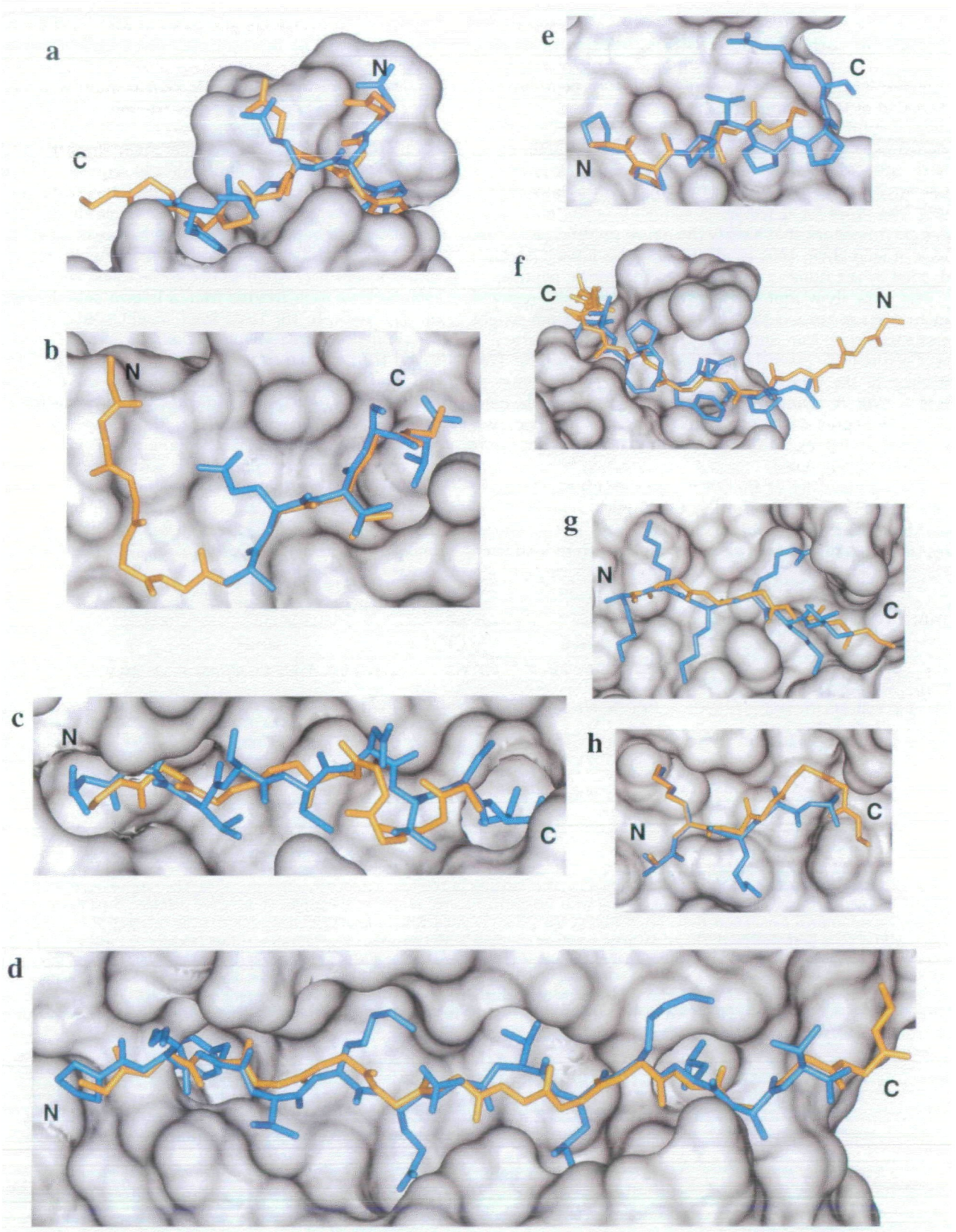

Fig. 7. **Views of the complexes of designed peptide (yellow) and the target protein (white surface).** The known-peptides are shown in blue. The designed-peptides are the tops in final generation (Table IV). The peptides were designed from sought seeds, except for that of SH3 domain and smaller sites of Kap α, in which only breeding from known seeds was successful. Side chains are shown only if the side chains of designed peptides are identical to those of the known peptides. a: CheR. b: PDZ domain. c: MHC I. d: MHC II. e: SH3 domain. f: HIV1 protease. g: Larger NLS-binding site of Kap α. h: Smaller NLS-binding site of Kap α.

Fig. 6 along with plots of average scores for the top half of peptides in each generation with respect to generation number. The progress was monitored using two criteria: detection of main chain trace of the known peptides, and detection of their sequences. The scoring system for detection of main chain trace is defined as the number of designed peptide residues that have an rmsd (of main chain N, C_α, and C atoms) of less than 1.0 Å from the known peptide, divided by the number of residues in the known peptide. The score for detection of sequence is the number of designed residues that satisfy the above condition and that have a side chain that is identical to the known peptide, divided by the number of residues in the known peptide.

The plots show that the known seeds generally resulted in an average score better than those obtained from sought seeds (compare thin and thick plots in Fig. 6). Accuracy in seed structure is essential for peptide-breeding results. This is clear in the cases of SH3 domain and smaller sites of Kap α (Fig. 6, e and h), in which the known seeds have grown to peptides that are similar to the known peptides, to a certain extent, while the breedings from sought seeds have completely failed.

Main chain trace of the known peptides was efficiently detected (thin-solid plots in Fig. 6 and stick-models in Fig. 7). More than 60% of the main chain trace of the known peptides was reproduced when the peptides were bred from known seeds, except for the cases of MHC I (Fig. 6c) and SH3 domain (Fig. 6e). Although the results from sought seeds are generally worse than those from known seeds, comparable results for CheR, PDZ domain and MHC II were given by both seed types (compare thick-solid and thin-solid lines in Fig. 6, a, b, and d).

Compared to the efficiency for main chain trace detection, however, detection of amino acid sequence of peptides was less efficient (compare solid and broken plots in Fig. 6). This is apparent from the fact that no case showed a significant increase in the score for the sequence during breeding from known seeds (thin-broken lines in Fig. 6). Each of the initial scores was equal to one correct side chain per peptide, because each residue from a known peptide was used for the seeds. In the tests from sought seeds, the score of the sequence increased slightly, except for the cases of Kap α and SH3 domain (thick-broken lines in Fig. 6). However, the final scores are equal to or less than the initial scores from known seeds, which indicates that each designed peptide has only one correct side chain on average. The peptides obtained from the final populations as the top (which has the highest E_{tot} value) or the best (which is the closest to the known peptide) peptide are shown in Table IV and Fig. 7. The top peptides contain up to two correct side chains, and the best peptides have up to four correct side chains, indicating that a complete set of correct side chains

TABLE IV. **Summary of peptide-breeding tests.**

Protein	CheR	PDZ domain	MHC I	MHC II
Known peptide[a]	N**W**ET**F**	KQTS**V**	T**L**TS CNT S**V**	P K**Y** VKQNTL K**L**AT
Designed peptide[b]				
Top from known seed	TPDL**W**I I W	VV VYDWPI TTF	D**L**W**L** I PWP	EWTWVCI YNI Y**LA**WP
Best from known seed	AALL**W**E**V**M	L TYI W**K**I TTH	T**L**SN PW	EWTWVCI YNI Y**LA**WP
Top from sought seed	W**D**W**E**F**E** FS W	WVFQF STTL H	WPTG YPWDD	WS W**I** F**I** I FG**F** TK**F** QK
Best from sought seed	GV**W**V**E**W YAW	EW**K**I T**F**V	T**L**RWVAW	P**KF** VYEGETW**L**DP
Cluster analysis of designed residues[c]				
Detected residues	. **WE**. W	. . TTL	. **L**.	**P** . **F** **V**. E. . . . **L**. .
Source of seeds	Sought	Sought	Sought	Sought
Total number of clusters	18	28	18	38
Trace detection	6	17	5	35
Sequence detection	2	1	1	3

[a]Sequence of peptides co-crystalized with the proteins. The residues in boldface letters are consensus sites. The residues indicated by asterisks were used for the seeds in peptide modeling with InsightII/Discover program. [b]Examples of designed peptides. Tops have the highest fitness values in final population. Bests are closest to the known peptide. Known or sought seed indicates the category of seeds from which the peptides have been designed. The underlined residues are close to the corresponding known residues (rmsd{C, C_α, and N} < 1.0 Å). The residues in boldface letters are identical to that of the known residues. [c]Results of cluster analysis. Items: Detected residues, the consensus side chains that are detected by the system; Source, category of seeds from which the presented results were obtained; No. clusters, number of designed-residue clusters; Trace detection, the numbers of clusters that are close to the known-residues (rmsd{C, C_α, and N} < 1.0 Å); Sequence detection, the number of clusters that are close to the known residues and have identical side chains to the known residues.

TABLE IV. **Summary of peptide-breeding tests (continued).**

Protein	SH3 domain	HIV1 protease	Kap α larger site	Kap α smaller site
Known peptide	P**P**P**V**P**P**R	SL N**F**P**I** V	K**K**K**R**K**V**	A K**K**AA
Designed peptide				
Top from known seed	I **PV**YP	VYWI **F**V**F**Y LI YP	NWYTAP **KA**T**KT**	**WK**F P WD W
Best from known seed	I **P**P**L**P	DL**Y**W**I** **F**I **I**V QP	I D**KKRQ**V**L**	A **W**K**F**A P K
Top from sought seed	P**E**L**V**PP	LVI F**Y** I I I D**T** YSVY AW	**W**Y**V**Y**I** V**Y**	A **H**P**H**F I
Best from sought seed		TY**YWNF**CDWLVKWP	G**F**V**F**I HC	
Cluster analysis of designed residues				
Detected residues	**PPP**. **PP**.	. . . **NF**. **I** .	. **K**. . **KV**	. . . . .
Source of seeds	Known	Sought	Known	Known
Total number of clusters	15	44	21	12
Trace detection	11	12	11	3
Sequence detection	5	3	3	0

was not detected in the peptide-breeding process.

The difficulty of sequence detection might be due to a limit in the searchable sequence space. The maximum variety of peptides (combinations of sequence and rotamer) that can be tested in one peptide-breeding run is in the order of 10^4. This number is roughly equal to the complete sequence variety of tripeptide ($20^3 = 0.8 \times 10^4$). When rotamer variation is taken into account, this number is reduced by ~10^2-fold. Apparently, only an extremely limited area in sequence-conformation space can be sought in peptide-breeding.

Key Residue Detection in Peptide-Breeding—The difficulty of sequence detection prompted an elaborate analysis in order to derive significant sequence information from the designed peptides. Accordingly, the residues of the designed peptides were subjected to a clustering analysis to find particular residues that appeared frequently in the final population.

The designed peptides from final generations were divided into residues, and the residues that showed an rmsd of main chain atoms (N, C_α, and C) of less than 1.0 Å and had the same side chain were combined into the same cluster. Clusters composed of more than 150 members (10% of the maximum 1,500) were selected and compared with the consensus sequence of the known-peptides. From 13 to 44 clusters were found for the proteins as a result of the clustering analysis (Table IV). The comparison showed that key residues of the peptide–protein interaction were indeed detected in the peptide-breeding process. The following discussion will be based on the designed peptides from sought seeds, unless otherwise mentioned.

CheR binds to the C-terminal peptide of the chemotaxic receptor protein. The consensus sequence of the receptor's termini is W/FXXF/– (*29*), which confers the two key residues in this interaction. Among the 18 clusters of residues from the designed-peptide, a cluster of Trp designed residues was found close to the N-terminal consensus Trp of the known peptide (Table IV). Another cluster of Trp, which is similar to the C-terminal consensus Phe, was also found.

PDZ domain binds to the C-terminal peptide of proteins that have consensus sequence S/TXV/L (*30*). At the Ser/Thr consensus site, a cluster of Thr residues was found. At the C-terminal consensus Val/Leu position, a cluster of Leu residues was found.

HIV1 protease processes viral polypeptide, and prefers Phe/Leu at the N-terminal side and Pro at the C-terminal side to the scissile bond (*34*). A cluster of Phe residues was found at the consensus Phe position. However, no cluster was found at the C-terminal Pro site.

MHC I and II molecules bind to a variety of antigen peptides in order to present these molecules to T-cell receptors. Two anchor residues of the peptide are known to flank the sequence that is to be presented (*31*, *32*). For MHC II, the N- and C-terminal anchors are known to be Tyr and Leu, respectively. A cluster of Phe, similar to the consensus Tyr, was found at the N-terminal anchor site, and a cluster of Leu was found at the C-terminal anchor site. Although a cluster of Leu residues was found in the position of the N-terminal anchor Leu/Ile of MHC I, no cluster for the C-terminal anchor Val/Leu was observed because a majority of the designed peptides did not reach this position.

Sequences of peptide have been reported that were screened from phage-display libraries with the same MHC II molecule used in this study (*36*). The 60 nonapeptide sequences from the screened library were compared with non-redundant 163 nonapeptide sequences designed with EmPLiCS from sought seeds. As the result, about 13% (22 of total 163 peptides) of designed peptides showed maximum 4 matches in 9 residues (Fig. 8a). The frequency is higher than the expected value by 7.8 standard deviations. The detected residues are mainly localized to the positions 1, 4, and 9 of the sequences (Fig. 8b); these positions are known to define the specific motif of the MHC II–binding peptides (*40*). Since the majority of designed peptides have a main chain structure close to that of the known peptide (Fig. 7d), most of the residue clusters can be assigned to one of the sequence positions. Among the 25 residue clusters belonging to the main chain trace, 18 were amino acids that have been found at the corresponding positions of the peptides from the screened library (Fig. 8c). The comparison shows that the designed sequences are significantly inclined to that of the high-affinity peptides of MHC II molecule, and the key residues are detected more efficiently than others.

No consensus site for SH3 domain or Kap α was found among the peptides that were designed from sought seeds.

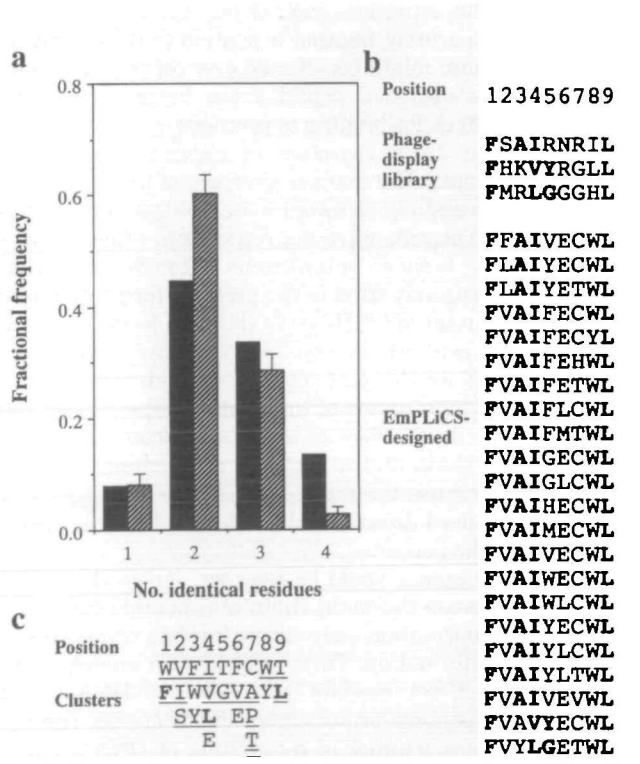

Fig. 8. **Comparisons of the EmPLiCS-designed peptides for MHC II with the peptides screened from phage-display libraries.** a: Histogram of fractional frequency of the designed peptides against the maximum number of identical residues with the peptides from the screened library. The filled bars show the frequencies of EmPLiCS-designed peptides. The hatched bars with error handle show the distribution of randomized peptides. b: Sequences of peptide from the screened library and the EmPLiCS-designed peptide that show 4/9 identity with the library peptides. Identical residues are shown in boldface letters. c: The residue clusters that belong to the major main chain trace of the designed-peptides. The underlined clusters were observed at the corresponding sites of the library peptides, and those in boldface letters were observed in 25% or more of the library peptides.

However, the system found consensus sites in the clusters that were designed from known seeds. SH3 domain is an adaptor molecule in the signal transduction pathway, and binds proline-rich peptide with consensus PXXPXR (*33*). Pro clusters of designed residues were found at the first and the second consensus Pro sites (Table IV). The residues that were directly inherited from the known seeds were excluded in this analysis in order to guarantee that the two sites were detected in the course of breeding.

Kap α is a carrier protein that binds to the nuclear localization signal (NLS). The NLS has several positively charged residues, although no strict consensus sequence has been observed (*35*). In the larger site of the protein, two clusters of Lys were detected (Table IV), although no positively charged residues were found for the smaller site even in the peptides from known seeds.

The results showed that many of the key residues were detected in peptide-breeding without detecting a complete sequence of the specific peptide. This implies that some of the key residues can be recognized out of the sequence context. This may be because the most of the known peptides bind to the proteins in extended conformation (Fig. 7). An extended conformation reduces the intrapeptide interaction, and consequently reduces interdependency among the residues. In this situation, each of the key residues would be detected separately, because a peptide that has only one correct side chain might be selected over others. The results show that the designed peptides can be used to predict some key residues for binding to proteins.

Main Chain Trace Detection in Peptide Breeding—As already mentioned, the main chain trace of the known peptides was successfully detected even though complete sequences of the peptide were not reproduced (Figs. 6 and 7). This is partly because interactions through main chain atoms are extensively used in the proteins for peptide binding. Among a total of 68 H-bonds that are formed between proteins and peptides in crystal structures of the seven complexes used for this test, 52 (77% of total) were found to involve main chain atoms of the peptides. The largest fraction of the bonds, 36 (53% of total), are formed between a protein's side chain and a peptide's main chain. This was unexpected, because the main chain atoms of peptide were not able to be used directly to discriminate between specific and non-specific peptides.

Main chain atoms would be, however, indirectly used for specificity, because the main chain of a peptide can take a particular conformation only when its side chains do not hinder the conformation. There might be an analogy of the direct and indirect readout strategies in DNA–protein interaction in peptide–protein interaction (*41–43*). The indirect readout is a manner of recognition of DNA sequence with a sparse direct contact between bases and the protein molecule. The specificity is thought to be generated from a difference in free-energy cost for the change in DNA structure among different nucleotide sequences.

It is possible that some peptide-binding proteins cast peptides into a specific main chain conformation and observe how well the side chains of peptides fit into the specific main chain conformation. Eighteen out of the 28 sample peptides of EF^2 (Table III) form more than two-thirds of the total number of H-bonds through their main chain atoms, and only 6 of these form the majority of H-bonds through side chain atoms, suggesting that recognition through peptide's main chain is a widely accepted strategy for peptide-protein interaction. Since the strategy requires extensive interaction of the main chain, the main chain trace might be readily detected for such proteins.

The results of the present study suggest that when a similar main chain trace is frequently observed among designed peptides, the trace might be a requirement of peptide-binding protein, even if the side chains of the peptides are found to be varied among these peptides. In addition, the newly developed system can be used for predicting the framework of interaction between proteins and their specific peptide ligands.

Significance of the Performance of EmPLiCS—Known peptides were also modeled by using InsightII/Discover program to evaluate the results from EmPLiCS system. The modeling was performed as detailed in the "MATERIALS AND METHODS" section. Each peptide was modeled twice from two different known seeds, indicated by asterisks in Table IV, and the better scores of main chain trace detection are shown in Fig. 6 in comparison with the scores of EmPLiCS-designed peptides.

The final scores of main chain trace detection of EmPLiCS-designed peptides (known seed design, thin-solid plots in Fig. 6) are always equal to or better than that of the InsightII/Discover-designed peptides (gray-horizontal lines in Fig. 6). It should be emphasized that even if the peptides were modeled from sought seeds, EmPLiCS showed a comparative performance in more than half of the cases, namely, CheR, PDZ domain, MHC II, HIV1 protease, and larger NLS-binding site of Kap α (compare thick-solid plots and gray-horizontal lines in Fig. 6, a, b, d, f, and g). Considering the complete sequences were given to InsightII/Discover modeling, while no sequence information was given to EmPLiCS, the performance of EmPLiCS was significantly better. The result shows that the main chain trace of a peptide ligand can not be easily detected even if the peptide-bound structure of protein and the sequence of peptide are known, and that the usage of the empirical rules made a significant improvement in detection efficiency.

Conclusion—The performance of the empirical peptide-ligand prediction system, EmPLiCS, was described. The performance of the system on several known peptide-protein complexes can be summarized into three points that suggest applications and potential directions for further improvement of the system. First, accuracy in the initial structure of peptides (seeds) is the most critical for design. It suggests finer steps in translation, rotation and rotamers in the seed-finding process, even though this will require more computational time. Second, sequence detection is rather inefficient, possibly due to the larger sequence space, compared to that which is scannable using the system. Nonetheless, the system detected some key residues according to the cluster analysis of the designed residues. Because residue clusters are frequently found on a continuous main chain trace, the clusters can be assigned to one of the sequence positions. This information might be used for experimental methods, such as combinatorial synthesis, in which the positional information helps to reduce the size of the library (*44*). Third, the system detected the main chain trace without reproducing complete sequences of specific peptides. The information on the preferred main chain structures can be used as a scaffold for further improve-

ment of designed peptides. This suggests a strategy in which the system is scheduled to seek for the appropriate main chain trace first, and side chain structures are examined after the main chain structures are fixed. The information of main chain trace can be used to deduce the residues of protein that interact with peptide ligand. Furthermore, the ability of main chain trace detection might be used in designing or predicting protein–protein interactions, which is also the important target of empirical methods in structural genomics.

This work was done at the computer-aided design facility of the Venture Business Laboratory of Nagoya University.

REFERENCES

1. Wlodawer, A. and Erickson, J.W. (1993) Structure-based inhibitors of HIV-1 protease. *Annu. Rev. Biochem.* **62**, 543–585
2. Schevitz, R.W., Bach, N.J., Carlson, D.G., Chirgadze, N.Y., Clawson, D.K., Dillard, R.D., Draheim, S.E., Hartley, L.W., Jones, N.D., Mihelich, E.D., Olkowski, J.L., Snyder, D.W., Sommers, C., and Wery, J.-P. (1995) Structure-based design of the first potent and selective inhibitor of human non-pancreatic secretory phospholipase A_2. *Nat. Struct. Biol.* **2**, 458–465
3. Lam, P.Y.S., Jadhav, P.K., Eyermann, C.J., Hodge, C.N., Ru, Y., Bacheler, L.T., Meek, J.L., Otto, M.J., Rayner, M.M., Wong, Y.N., Chang, C.-H., Weber, P.C., Jackson, D.A., Sharpe, T.R., and Erickson-Viitanen, S. (1994) Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* **263**, 380–384
4. Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D., and Perry, K.M. (1993) Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259**, 1445–1450
5. Kuntz, I.D. (1992) Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082
6. Amzel, L.M. (1998) Structure-based drug design. *Curr. Opin. Biotech.* **9**, 366–369
7. Wade, R.C. (1997) 'Flu' and structure-based drug design. *Structure* **5**, 1139–1145
8. von Itzstein, M., Wu, W.-Y., Kok, G.B., Pegg, M.S., Dyason, J.C., Jin, B., Phan, T.V., Smythe, M.L., White, H.F., Oliver, S.W., Colman, P.M., Varghese, J.N., Ryan, D.M., Woods, J.M., Bethell, R.C., Hotham, V.J., Cameron, J.M., and Penn, C.R. (1993) Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363**, 418–423
9. Aronov, A.M., Verlinde, C.L.M.J., Hol, W.G.J., and Gelb, M.H. (1998) Selective tight binding inhibitors of trypanosomal glyceraldehyde-3-phosphate dehydrogenase via structure-based drug design. *J. Med. Chem.* **41**, 4790–4799
10. Babu, Y.S., Ealick, S.E., Bugg, C.E., Erion, M.D., Guida, W.C., Montgomery, J.A., and Secrist, III J.A. (1995) Structure-based design of inhibitors of purine nucleoside phosphorylase. *Acta Cryst.* **D51**, 529–535
11. Nienaber, V.L., Mersinger, L.J., and Kettner, C.A. (1996) Structure-based understanding of ligand affinity using human thrombin as a model system. *Biochemistry* **35**, 9690–9699
12. Merrifield, R.B. (1963) Solid phase peptide synthesis. I. The synthesis of a tetrapeptide. *J. Am. Chem. Soc.* **85**, 2149–2154
13. Norman, T.C., Smith, D.L., Sorger, P.K., Drees, B.L., O'Rourke, S.M., Hughes, T.R., Roberts, C.J., Friend, S.H., Fields, S., and Murray, A.W. (1999) Genetic selection of peptide inhibitors of biological pathways. *Science* **285**, 591–595
14. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., and Abola, E.E. (1998) Protein data bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Cryst.* **D54**, 1078–1084
15. Singh, J., Saldanha, J., and Thornton, J.M. (1991) A novel method for the modelling of peptide ligands to their receptors. *Protein Eng.* **4**, 251–261
16. Laskowski, R.A., Thornton, J.M., Humblet, C., and Singh, J. (1996) X-SITE: Use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. *J. Mol. Biol.* **259**, 175–201
17. Kono, H. and Sarai, A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins Struct. Funct. Genet.* **35**, 114–131
18. Moon, J.B. and Howe, W.J. (1991) Computer design of bioactive molecules: A method for receptor-based de novo ligand design. *Proteins Struct. Funct. Genet.* **11**, 314–328
19. Frenkel, D., Clark, D.E., Li, J., Murray, C.W., Robson, B., Waszkowycz, B., and Westhead, D.R. (1995) PRO_LIGAND: An approach to de novo molecular design. 4. Application to the design of peptides. *J. Comput.-Aided Mol. Design* **9**, 213–225
20. Murray, C.W., Clark, D.E., and Byrne, D.G. (1995) PRO_LIGAND: An approach to de novo molecular design. 6. Flexible fitting in the design of peptides. *J. Comput.-Aided Mol. Design* **9**, 381–395
21. Böhm, H.-J. (1996) Towards the automatic design of synthetically accessible protein ligands: Peptides, amides and peptidomimetics. *J. Comput.-Aided Mol. Design* **10**, 265–272
22. Schneider, G., Schrödl, W., Wallukat, G., Müller, J., Nissen, E., Rönspeck, W., Wrede, P., and Kunze, R. (1998) Peptide design by artificial neural networks and computer-based evolutionary search. *Proc. Natl. Acad. Sci. USA* **95**, 12179–12184
23. Sippl, M.J. (1990) Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**, 859–883
24. Bowie, J.U., Lüthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170
25. Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992) A new approach to protein fold recognition. *Nature* **358**, 86–89
26. Sánchez, R. and Sali, A. (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* **95**, 13597–13602
27. Rychlewski, L., Zhang, B., and Godzik, A. (1999) Functional insights from structural predictions: analysis of the *Escherichia coli* genome. *Protein Sci.* **8**, 614–624
28. Zhang, B., Rychlewski, L., Pawlowski, K., Fetrow, J.S., Skolnick, J., and Godzik, A. (1999) From fold predictions to function predictions: Automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104–1115
29. Djordjevic, S. and Stock, A.M. (1998) Chemotaxis receptor recognition by protein methyltransferase CheR. *Nat. Struct. Biol.* **5**, 446–450
30. Doyle, D.A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell* **85**, 1067–1076
31. Madden, D.R., Garboczi, D.N., and Wiley, D.C. (1993) The antigenic identity of peptide-MHC complexes: A comparison of the conformations of five viral peptides presented by HLS-A2. *Cell* **75**, 693–708
32. Stern, L.J., Brown, J.H., Jardetzky, T.S., Gorga, J.C., Urban, R.G., Strominger, J.L., and Wiley, D.C. (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* **368**, 215–221
33. Lim, W.A., Richards, F.M., and Fox, R.O. (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. *Nature* **372**, 375–379
34. Swain, A.L., Miller, M.M., Green, J., Rich, D.H., Schneider, J., Kent, S.B.H., and Wlodawer, A. (1990) X-ray crystallographic structure of a complex between a synthetic protease of human immunodeficiency virus 1 and a substrate-based hydroxyethylamine inhibitor. *Proc. Natl. Acad. Sci. USA* **87**, 8805–8809
35. Conti, E., Uy, M., Leighton, L., Blobel, G., and Kuriyan, J. (1998) Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin α. *Cell* **94**, 193–204
36. Hammer, J., Takacs, B., and Sinigaglia, F. (1992) Identification

of a motif for HLA-DR1 binding peptides using M13 display libraries. *J. Exp. Med.* **176**, 1007–1013
37. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784
38. Singh, J. and Thornton, J.M. (1985) The interaction between phenylalanine rings in proteins. *FEBS Lett.* **191**, 1–6
39. Dayhoff, M.O., Eck, R.V., and Park, C.M. (1972) A model of evolutionary change in proteins in *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.) Vol. 5, pp. 89–99, National Biomedical Research Foundation, Silver Spring, Washington, DC
40. Hammer, J., Valsasnini, P., Tolba, K., Bolin, D., Higelin, J., Takacs, B., and Sinigalia, F. (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell* **74**, 197–203
41. Koudelka, G.B., Harrison, S.C., and Ptashne, M. (1987) Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature* **326**, 886–888
42. Otwinowski, Z., Schevitz, R.W., Zhang, R.-G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **335**, 321–329
43. Martin, A.M., Sam, M.D., Reich, N.O., and Perona, J.J. (1999) Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nat. Struct. Biol.* **6**, 269–277
44. Furka, A., Sebestyen, F., Asgedom, M., and Dibo, G. (1991) General method for rapid synthesis of multicomponent peptide mixtures. *Int. J. Peptide Protein Res.* **37**, 487–493