12 Vilan, A. *et al.* (1998) Real-time electronic monitoring of adsorption kinetics: Evidence for two-site adsorption mechanism of dicarboxylic acids on GaAs(100). *J. Phys. Chem. B.* 102, S3307–S3309
13 Wu, D.G. *et al.* (2000) Novel NO biosensor based on the surface derivatization of GaAs by 'hinged' iron porphyrins. *Angew. Chem., Int. Ed. Engl.* 39, 4496–4500
14 Wu, D.G. *et al.* (2001) Direct detection of low concentration NO in physiological solutions by new GaAs-based sensor. *Chem. Eur. J.* 7, 1743–1749
15 Mandelis, A. and Chrisofides, C. (1993) *Physics, Chemistry, and Technology of Solid State Gas Sensor Devices*, John Wiley & Sons
16 Hagfeldt, A. and Gratzel, M. (2000) Molecular photovoltaics. *Acc. Chem. Res.* 33, 269–277
17 Miles, J.L. and McMahon, H.O. (1961) Use of monomolecular layers in evaporated-film tunneling devices. *J. Appl. Phys.* 32, 1176–1177
18 Polymeropoulos, E.E. and Sagiv, J. (1978) Electrical conduction through adsorbed monolayers. *J. Chem. Phys.* 69, 1836–1847
19 Ulman, A. (1991) *An Introduction to Ultrathin Organic Films*, Academic Press
20 Tredgold, R. *et al.* (1984) Structural effects on the electrical conductivity of Langmuir-Blodgett multilayers of cadmium stearate. *J. Phys. D* 17, L5–L8
21 Vilan, A. *et al.* (2000) Molecular control over Au/GaAs diodes. *Nature* 404, 166–168
22 Selzer, Y. and Cahen, D. (2001) Fine tuning of Au/SiO_2/Si diodes by varying interfacial dipoles using molecular monolayers. *Adv. Mater.* 13, 508–511
23 Kralchevsky, P.A. and Nagayama, K. (2000) Capillary interactions between particles bound to interfaces, liquid films and biomembranes. *Adv. Colloid Interface Sci.* 85, 145–192
24 Slowinski, K. *et al.* (1999) Mercury–mercury tunneling junctions: 1. Electron tunneling across symmetric and asymmetric alkanthiolate bilayers. *J. Am. Chem. Soc.* 121, 7257–7261
25 Holmlin, R.E. *et al.* (2001) Correlating electron transport and molecular structure in organic thin films. *Angew. Chem., Int. Ed. Engl.* 40, 2316–2320
26 Krüger, J. *et al.* (2000) Controlling electronic properties of TiO_2 by adsorption of carboxylic acid derivatives. *Adv. Mater.* 12, 447–451
27 Rhoderick, E.A. and Williams, R.H. (1988) *Metal-semiconductor Contacts*, Clarendon Press
28 Tung, R.T. (1992) Electron transport at metal-semiconductor interfaces: General theory. *Phys. Rev. B* 45, S13509–S13523
29 van Ruyven, L.J. (1964) The position of the Fermi Level at a heterojunction interface. *Phys. Stat. Sol.* 5, K109
30 Tredgold, R. and Badawy, Z.E. (1985) Increase of the Schottky barrier height at GaAs surfaces by carboxylic acid monolayers and multilayers. *J. Phys. D* 18, 103–109
31 Tredgold, R. and Badawy, Z.E. (1985) GaAs Schottky diodes incorporating Langmuir-Blodgett layers of pre-formed polymers. *J. Phys. D* 18, 2483–2487
32 Gal, D. *et al.* (1997) Engineering the interface energetics of solar cells by grafting molecular properties onto semiconductors. *Proc. Ind. Acad. Sci. Chem. Sci.* 109, 487–496
33 Zuppiroli, L. *et al.* (1999) Self-assembled monolayers as interfaces for organic opto-electronic devices. *Euro. Phys. J. B.* 11, 505–512
34 Campbell, L.H. *et al.* (1997) Controlling charge injection in organic electronic devices using self-assembled monolayers. *Appl. Phys. Lett.* 71, 3528–3530
35 Chai, L. and Cahen, D. Electric signal transfer through nm-thick molecular bilayers. *Mater. Sci. Eng. C.* (in press)
36 Bruening, M. *et al.* (1994) Polar ligand adsorption controls semiconductor surface potentials. *J. Am. Chem. Soc.* 116, 2972–2977
37 Bruening, M. *et al.* (1995) Controlling the work function of CdSe by chemisorption of benzoic acid derivatives and chemical etching. *J. Phys. Chem.* 99, 8368–8373
38 Bastide, S. *et al.* (1997) Controlling the work function of GaAs by chemisorption of benzoic acid derivatives. *J. Phys. Chem.* 101, 2678–2684
39 Bruening, M. *et al.* (1997) Simultaneous control of surface potential and wetting of solids with chemisorbed multifunctional ligands. *J. Am. Chem. Soc.* 119, 5720–5728
40 McClellan, A. (1963) *Tables of Experimental Dipole Moments*, WH Freeman
41 Pople, J.A. (1999) Quantum chemical models (Nobel lecture). *Angew. Chem., Int. Ed. Engl.* 38, 1894–1902

Protein therapeutics: promises and challenges for the 21st century

Zhiping Weng and Charles DeLisi

Recent advances in massively parallel experimental and computational technologies are leading to radically new approaches to the early phases of the drug production pipeline. The revolution in DNA microarray technologies and the imminent emergence of its analogue for proteins, along with machine learning algorithms, promise rapid acceleration in the identification of potential drug targets, and in high-throughput screens for subpopulation-specific toxicity. Similarly, advances in structural genomics in conjunction with *in vitro* and *in silico* evolutionary methods will rapidly accelerate the number of lead drug candidates and substantially augment their target specificity. Taken collectively, these advances will usher in an era of predictive medicine, which will move medical practice from reactive therapy after disease onset, to proactive prevention.

Substantial improvement is needed in almost every aspect of drug development. Scientists' ability to design drugs that exploit disease phenotype is limited, knowledge about the range of actions of even common medications is meager and the ability to customize drugs specific for human subpopulations remains beyond reach. In addition, the production pipeline is long and expensive, and the effectiveness of its end products – some 5000 drugs currently in commerce, targeted at ~500 medical conditions – varies with disease target and with individual genotype in ways that are poorly understood.

This picture, however, is about to change dramatically and there is excellent reason to be sanguine about the future because biology has begun to emerge as a predictive, quantitative science with a rational basis for rapid design and discovery. As a consequence, the coming decades will witness a radical increase in the number of available drugs and targets and in their clinical efficacy and safety, and a reduction in their cost of production and time to market.

The new era that is now upon us is defined and symbolized by the genomic revolution – a revolution

 0167-7799/02/$ – see front matter PII: S0167-7799(01)01846-7

that encompasses high-throughput sequencing, and the technologies that exploit and add to the information it generates. The latter include various DNA microarray methods that monitor genomic change [1,2], emerging array methods for monitoring protein profiles [3,4], high-throughput methods for genetic characterization of disease diathesis and resistance [5], high-performance computing for biological discovery [6] and real-time data analysis and data integration.

These technologies, as indicated below, will accelerate the discovery and molecular characterization of disease-specific genetic pathways [7,8], and of ligand-specific toxicity and metabolic pathways [9]. They will thereby accelerate the discovery of protein therapeutics and the identification of candidate drug targets, while moving us toward individualized medicine (therapy that takes into account genetic markers for predisposition to drug side effects and/or efficacy). In this article, we focus predominantly on a conceptual overview of genomic-related methods for identifying lead candidates for targets and for drug design. We specifically omit the important area of lead refinement and soft drug design [10], which takes into account the distribution and metabolism of the drug in a physiological system, and its activity at the intended site of action. We further restrict ourselves to protein therapeutics, omitting the important and currently more successful arena of small-molecule drugs [11,12].

Microarrays and the challenge of diversity

The development of DNA microarrays to monitor the expression of all or a substantial fraction of the expressed genes of a cell [13,14], and more generally for characterizing the change in genomic expression that accompanies normal development and disease progression, will have deep ramifications for identifying highly specific disease targets and for customizing drugs [7,8,15,16]. Microarray technologies vary greatly, each has its own set of merits and limitations. They all involve immobilizing DNA probes on a glass or nylon surface. The probes are reverse complements of target regions on mRNA (or cDNA) whose concentration, or expression level, they monitor through hybridization. They can be manufactured either as PCR products of intact cDNA (300–1000 bases long) spotted onto the surface [1], or as short oligonucleotides (20–30 bases long) synthesized *in situ* [2]. The latter requires only the sequences of the target genes (not the DNA itself), and thereby maximally exploits the reference human genome sequence, which is becoming increasingly informative as the amount of sequence and fold coverage (how many folds clone libraries cover a genome) increase. A 1–2 cm^2 array can, in principle, probe for several hundred thousand genes. Moreover, when oligonucleotide probes are much shorter than their targets, they can be selected to optimize specificity.

Zhiping Weng*
Charles DeLisi*
Biomedical Engineering Dept and Bioinformatics Program, Boston University, Boston MA 02215, USA.
*e-mail: zhiping@bu.edus

Applications of microarray technology to disease biology and diagnosis, and especially to cancer, have been extensive. The concept that large numbers of transcript profiles might permit stratification of previously unrecognized cancer subtypes is particularly important [17]. This was shown clinically when Golub *et al.* [18] reported that transcript profiles of leukemia cells could be divided into acute myeloid and acute lymphoblastic subtypes without the researcher knowing on what basis they could be distinguished. Alizadeh *et al.* [19] subsequently identified gene signature profiles for two subgroups of distinctly differentiated B cells, which correlated with patient survival. Alon *et al.* [20] studied gene expression in samples of tumor colon tissue and were able to distinguish them from normal colon tissue samples on the basis of gene expression. Examples of similar studies include breast cancer [21] and ovarian cancer [22].

An increasingly common practice is to use customized arrays to probe for particular sets of genes, for example, if oxidant damage is hypothesized to be important in the etiology of lung cancer [23], a stress array consisting of probes for perhaps hundreds of antioxidant and repair genes could be used; if polymorphism in apoptosis genes is believed to underlie resistance and susceptibility to tuberculosis, an array to probe for expression of apoptosis genes can be designed to characterize gene expression in infected alveolar macrophages [24]. In this way, the high-throughput advantages of arrays for discovery are combined with hypothesis-driven research.

In addition to showing that expression patterns cluster to form diagnostic fingerprints, these and other results offer promise for identifying genes and pathways that are potential therapeutic targets. Thus Golub *et al.* [18] found an overexpression of the HOXA9 oncogene associated with refractoriness to therapy directed against acute myelogenous leukemia, and Alizadeh *et al.* uncovered sets of genes implicated in apoptosis inhibition [19].

The upregulation or downregulation of genes can either be the cause or the result of the disease. Moreover, owing to the complexity of gene regulation and the multigenic nature of most diseases, microarray experiments on disease tissues typically uncover hundreds of genes with altered expression. Thus, to pinpoint specific genes as drug targets, traditional methods are still required, although they are increasingly being integrated with genomic techniques [25,26]. What microarrays do best is high-throughput screening – and this can be for target validation as well as for identification. For example, if a gene deletion or mutation produces a genomic expression pattern similar to that of a disease, the gene and its product are potential drug targets. Similarly, microarrays can be used to test the efficacy and toxicity of drug candidates by selecting those that can best recover

the normal pattern of gene expression. These experiments can be performed on cells with different genotypes to predict functional variation in the response of individuals to different drugs. Such studies over time, when coupled with advances in protein technologies, will provide the extensive and diverse data-structures and the profound understanding of cell biology, required for a truly predictive medicine.

Corresponding technologies for monitoring changes in protein abundance are not available, but the field of proteomics is active and growing rapidly [3,4,27–29]. Proteome technologies are important because most current drug targets are proteins, but also because of the variable and unreliable correlation between gene and protein expression, and in the post-translational protein modifications [30] that are responsible for realizing the signaling and information processing that regulate cell behavior.

A particularly relevant example of the latter is modification of Ras by the addition of farnesyl hydrocarbon. It is well known that hyperactivity of the *ras* gene and consequent activity of the Ras pathway [15] is implicated in several cancers, perhaps the best studied being colorectal cancer. This suggests farnesyl transferase inhibitors as a lead therapeutic; indeed several are in clinical trials. More generally, because other proteins in the Ras pathway – for example mitogen activated protein kinase and mitogen activated protein kinase kinase– are also post-translationally modified, we can reasonably expect an effective proteomics strategy to multiply initial leads by many fold.

Identification of natural ligands

Once a protein is deemed to be a drug target, its biological ligand logically becomes a drug candidate. The availability of all full-length cDNAs in the human genome has made it possible for three proteomic methods to be developed to identify binding partners [27]. They are briefly described below.

Mass spectrometry

Mass spectrometry can be used to detect the ligand directly. The receptor molecule is first immobilized on a bead and then treated with cell lysate. After the nonspecific binders are washed away, the complex is eluted. Proteolysis of the purified complex leads to peptide fragments, whose masses can be determined precisely using mass spectrometry [31]. Knowing the masses of the proteolysis fragments is usually not sufficient to identify the ligand. However, the theoretical fragmentation spectra of all possible proteins can be compared with the observed spectrum to identify the mostly probable sequence [32]. The sensitivity of mass spectrometry allows this method to detect multi-component complexes such as the yeast nuclear-pore complex [33], the chloroplast of pea [34] and the interchromatin granule cluster [35].

Yeast two-hybrid system

A high-throughput biological alternative to the above method is the yeast two-hybrid system, which takes advantage of our knowledge of transcription machinery. The method involves fusions in two different yeast strains. In one strain a reporter gene is fused to the DNA binding domain of GAL4 (a protein that is a transcriptional activator) and in the other a cDNA library is fused to the GAL4 activation domain. When the two strains are mated, the reporter gene will be expressed only when the two GAL4 domains are in close proximity. The mated strains are grown under conditions that require the protein product of the reporter gene, and the surviving yeast cells are harvested and the ligand sequence uncovered by sequencing the inserts. This method can achieve very high throughput. Each of the GAL4 domains can be cloned with a cDNA library to create a protein–protein interaction map of the cell [36–39].

Display technologies

A third experimental approach to the identification of natural ligands is based on display technologies [40–42]. The traditional use of display technologies for *in vitro* selection is discussed in the next section. For *in vitro* selection, the display library is composed of combinatorially generated mutants of the molecule to be optimized. For identifying natural ligands however, the entire proteome is used to construct the display library.

The Darwinian theme *in vitro*

Display technologies are a family of experimental methods that permit combinatorial generation of diversity followed by selection and amplification of those molecules with the desired property; for example, tight binding to a receptor. One of the most important characteristics of display technologies is the ability to associate every protein with its genetic material (RNA or DNA). The protein is the 'displayed entity' of which function can be screened for. Although amino acid sequences of low abundance proteins cannot be readily determined, there are many high-throughput techniques for DNA sequencing and amplification. Thus, 'linking' every protein molecule to its oligonucleotide allows for the rapid 'decoding' of desirable proteins once they have been selected from the library. Two other important features of display technologies include the diversity and quality of the library and the screening strategy. The randomization and selection process can be iterated many rounds, accelerating the evolutionary process by nearly a billionfold.

The interplay between proteins and their genes can take on many forms, leading to different kinds of display technologies. Bacteriophage-based methods were the first to be developed, and remain the most widely used [43,44]. The oligonucleotide encoding the target protein is fused with the gene of a phage coat protein. By fusing oligonucleotides containing

random mutations with the phage gene, a library of phages each carrying a distinct peptide sequence as part of its coat protein can be produced. Thus the target protein is linked to its DNA through a phage particle. The phage library is added to a dish coated with the receptor molecule and unbound phages are washed away. Binding phages are harvested, amplified and sequenced to uncover the mutations that cause the improved binding. Ribosomal display involves the direct attachment of the target protein to its mRNA molecule through ribosome [45]. The RNA–peptide fusion technique covalently bonds a protein to its mRNA [46]. Other examples include flagellum display [47], yeast display [48] and mammalian cell-based display [49]. Several recent reviews provide excellent summaries of the different formats of display technologies as well as their strengths and weaknesses [44, 50–52].

There are practical limits to the size of display libraries – phage and cell-based displays are unable to handle more than 10^{11} library members and cell-free ribosomal displays can handle no more than 10^{14}. Direct synthesis of random oligonucleotides is easy to do but it is not the most efficient way to search the sequence space. Complete randomization of 24 nucleotide positions leads to 4^{24} distinct library members, or a 10^{14} library size. This only corresponds to eight amino acid positions. Of course, redundant codons and stop codons can be eliminated. This is achieved by mixing 20 presynthesized trinucleotide phosphoramidites [53], each corresponding to a different amino acid. Cho *et al.* used mRNA display to filter out frame shifts and stop codons in the randomized region prior to ribosomal display [54]. At perfect accuracy, the theoretical limit of library size thus corresponds to the complete randomization of 11 amino acid positions, which is a very small number compared with the average protein size (~150 amino acids); even binding sites typically contain 20–40 residues.

It is apparent from the above simple analysis that the library must be designed strategically. One option is to focus on positions of the molecule that are likely to improve functionality, such as at the binding site. Three-dimensional structural information can frequently guide the design. This is the case for antibodies and T-cell receptors, in which the binding site of six loops is clearly defined. The technique has produced an impressive array of antibodies whose the affinities for antigen far exceed those observed in nature [43].

Sometimes it is not obvious from the 3D structure which positions of the molecule are important for its function. Moreover, positions far away from the binding site can affect the function allosterically, or simply affect the stability of the protein. The extreme approach is to introduce random mutations throughout the gene using, for example, error-prone PCR or an *Escherichia coli* strain lacking DNA repair mechanisms (see [55] for different ways of generating sequence variety). Although the coverage on the sequence space is extremely low and most of the time unviable mutants are produced, it is a worthwhile avenue to explore when not much functional information is available. For example, it has been very difficult to display T-cell receptors, possibly owing to their low solubility or low stability. Kieke *et al.* successfully obtained T-cell receptor mutants that could be displayed on the surface of yeast from a random library produced using an *E. coli* mutator strain [56]. The library had only 6×10^7 members.

Stemmer has invented an ingenious technique for generating sequence diversity, called gene shuffling [57]. Related DNA sequences (e.g. all 20 copies of interferon α in the human genome) are amplified and randomly fragmented, and the fragments are reassembled using DNA polymerase in a self-priming fashion. The resulting chimeric molecules are selected using a display technology. Desirable progeny molecules are selected and can be bred again, accumulating multiple beneficial mutations.

One important feature of gene shuffling is its search strategy. Multi-parental recombination is an effective way to search through the sequence space, because the progenies that inherit beneficial mutations from multiple parents can be produced after crossover and can then be selected. Genetic algorithms do precisely that *in silico*. Computer simulations have shown that Genetic algorithms can find the maximum in a complex landscape rapidly [58].

Shuffling is an extremely efficient method for exploring the genetic diversity of natural sequences. The recent work by Chang *et al.* on evolving human interferon α is revealing [59]. The three most potent chimeras were more than 100-fold more active than the best parent. They were derived from five parental human genes but, strikingly, contained no random point mutations. This indicates that there is tremendous potential in the genetic diversity of natural sequences. Evolution quickly eliminates deleterious mutations, and would only tolerate moderately inferior mutations if they were compensated by beneficial ones. Thus, exploring natural mutations is analogous to searching through an infinitesimal sequence subspace that is largely free of deleterious mutations; in the meantime, beneficial mutations are enriched.

The challenge and promise of structural biology

Computational approaches to small-molecule drug discovery use geometric recognition algorithms to search small-molecule databases for structures complementary to target molecules [60,61], with the expectation that they will bind the target, and thereby modulate its activity [62]. This is an important start for both high-throughput screening and rational drug design. Although small molecules have an excellent track record for targeting enzymes

and ion channels, they are not as effective as proteins in blocking interactions between large macromolecules, such as occur in Ras or Src pathways. Also, non-biological small molecules tend to be substantially more toxic than human proteins. Proteins thus represent a growing class of therapeutic agents, in spite of difficulties related to their pharmacokinetic properties [63].

Protein–protein docking in the context of natural ligand identification involves searching a structural database by calculating the most stable complex that can form between each protein in the database and the potential target [64,65]. Predicting the stability of a pair is difficult partly because of the requirement for a rapidly evaluatable free-energy function with correctly balanced components (solvation, electrostatics, van der Waals or steric, and entropic effects), and partly because the free energy must be evaluated, for each pair of structures, over a large number of conformations. Including flexibility in surface structure means at a minimum allowing surface side-chain flexibility. Surface side-chains in the free molecules are likely to be in motion, rather than in a single dominant conformation. When the complex forms, those side-chains in the interface lock into a single favorable conformation. Consequently, not only must the rigid body rotational degrees of freedom of the molecules be searched but the side-chain conformations that provide the best interaction must also be found.

Early approaches used geometric criteria (steric) to match receptor and ligand starting from the bound conformation, and most of them assumed the location of the receptor binding site to be known. Recent developments attempt to dock unbound molecules, of which surface side-chains are not optimized for complex formation. Free energy functions that take account of solvation and electrostatics are now available [66,67], and more recently binding free energy functions taking into account solvation, electrostatics and entropy have been developed and validated [68–72].

Knowing the binding site of a molecule can substantially improve the performance of docking algorithms [73]. The finding that different ligands can bind to the same site on the receptor indicates that binding sites possess features that predispose them to ligand binding. For example, four natural proteins – protein A, protein G, neonatal Fc receptor and rheumatoid factor – all bind to the hinge region between the C_H2 and C_H3 domains on human immunoglobulin G. Random peptides also preferably bind to the same site [74]. Algorithms that predict binding sites can be based on structure [75,76] or sequence [77]. Experimental approaches have been developed to map binding sites by solving receptor structures in organic solvents [78,79].

Algorithms that do not assume known binding sites are also under development [80–83]. Notably, Fast Fourier Transform (FFT)-based methods, which search the entire 6D rotational and translational space of the ligand [73,80,84], perform relatively well in blind trials [85,86]. Gabb *et al.* applied a FFT algorithm with a steric-electrostatic-combined target function to ten unbound complexes, and ranked tens of near-correct structures in the top 4000 without knowing the binding site. Chen and Weng further developed the FFT algorithm with a target function that is tolerant to conformation changes. They performed a comprehensive study on 28 distinct protein–protein complexes. They ranked near-native ligand orientations in the top 2000 choices for 25 complexes. For three systems, their algorithm could identify the correct complex structure unambiguously.

Although there has been substantial progress in speed, current methods still require of the order of hours to dock a pair of proteins with a full 6D search, even with surface side-chain flexibility considered in an implicit way and with a simplified target function. As a result, all methods generate large numbers of false-positive structures, and thus, post-processing remains [66,69,87,88].

The computer power now available to relatively small research labs and the accessibility of Gig-Byte internal memory will continue to drive and accelerate progress in the foreseeable future. Individual labs now can afford dedicated Linux clusters with hundreds of nodes. Such clusters will enable the implementation of more complete target functions and the improved treatment of side-chain flexibility. The further development of empirical binding free energy functions will benefit from experimentally determined protein complex structures produced by structural genomics initiatives [89]. Very soon, the time to dock two proteins will be less than one hour. We can then expect that with 200 central processing units, an entire structural database with 30 000 members will be searched in less than one week. As computer speeds continue to increase and algorithms continue to improve, the time will continue to decrease, even allowing for increasing numbers of structures.

Protein design *in silico*

One of the goals of protein design is to achieve improved stability for a monomeric protein [90–92]. Thus, for example, Dahiyat and Mayo [93] began with the backbone of a 28-residue zinc finger and computationally screened 10^{27} sequences to find candidates with high stability. In particular they found computationally, and verified experimentally, compact structures that were more stable than the native zinc finger, and with no appreciable sequence similarity to known zinc fingers. More recently, Malakauskas and Mayo [94] reported a hyperthermophilic variant of the β1 domain of protein G, which has a melting temperature in excess of 100°C, maintains the fold and retains a significant level of binding to human IgG. The two

most important features of their method are the isomeric state approximation for side-chains (rotamers), and the implementation of the Dead-End Elimination theorem to efficiently search through the sequence space [95]. The success of Mayo's team has begun to stimulate additional research this area [96–100].

The same design principles used to increase the stability of a monomer can be applied to increase the stability and specificity of complexes. This problem is simpler than docking because the goal is to achieve improved stability to meet some design specification, and not necessarily to achieve the most stable complex. One would begin with the crystal structure of a complex that has low affinity for the ligand, and search for those sequences of the ligand that increases some combination of affinity and specificity. The design problem here requires searching through a vast sequence space, composed of all positions at the binding site. Compared with display technologies, which are limited to libraries with 10^{14} members, computational design can explore a much larger sequence space. Recent work by Looger and Hellinga [100] examined 2×10^{76} sequences in two days, distributed over eight 700 MHz Pentium III processors.

An example related to our own interests is the design of high affinity T-cell receptors for peptide – major histocompatibility complex (MHC) – complexes. The immune system is triggered when an infected cell displays on its surface, peptide fragments from the infective agent in association with host molecules encoded in the MHC. The peptide–MHC complex is recognized as foreign by cytotoxic T cells, which bind to and destroy the infected cell. T-cell receptors typically bind peptide–MHC complexes with low affinities, in the range of 10^{-5} M, relying on ancillary interactions after specificity is established, to increase stability [101,102]. Because the affinity of the native complex is low, and because this is an antigen–antibody-like system (which can achieve picomolar binding affinity [103,104]), it would be reasonable to expect to find sequences that increase the affinity by several orders of magnitude. This project is ongoing, but when an automated scheme is finally developed it could have substantial diagnostic and therapeutic applications.

Acknowledgement
Zhiping Weng is partially supported by NSF grant DBI#0078194.

References

1 Duggan, D.J. *et al.* (1999) Expression profiling using cDNA microarrays. *Nat. Genet.* 21, 10–14
2 Lipshutz, R.J. *et al.* (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* 21, 20–24
3 MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289, 1760–1763
4 Figeys, D. and Pinto, D. (2001) Proteomics on a chip. *Electrophoresis* 22, 208–218
5 Kruglyak, L. and Nickerson, D.A. (2001) Variation is the spice of life. *Nat. Genet.* 27, 234–236
6 Roos, D.S. (2001) Computational biology. Bioinformatics – trying to swim in a sea of data. *Science* 291, 1260–1261
7 Braxton, S. and Bedilion, T. (1998) The integration of microarray information in the drug development process. *Curr. Opin. Biotechnol.* 9, 643–649
8 Debouck, C. and Goodfellow, P.N. (1999) DNA microarrays in drug discovery and development. *Nat. Genet.* 21, 48–50
9 Nuwaysir, E.F. *et al.* (1999) Microarrays and toxicology: the advent of toxicogenomics. *Mol. Carcinog.* 24, 153–159
10 Bodor, N. and Buchwald, P. (2000) Soft drug design: general principles and recent applications. *Med. Res. Rev.* 20, 58–101
11 Archer, R. (1999) The drug discovery factory: an inevitable evolutionary consequence of high–throughput parallel processing. *Nat. Biotechnol.* 17, 834
12 Dove, A. (1999) Drug screening – beyond the bottleneck. *Nat. Biotechnol.* 17, 859–863
13 Fodor, S.P. *et al.* (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773
14 Schena, M. *et al.* (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16, 301–306
15 Garrett, M.D. and Workman, P. (1999) Discovering novel chemotherapeutic drugs for the third millennium. *Eur. J. Cancer* 35, 2010–2030
16 Freeman, T. (2000) High throughput gene expression screening: its emerging role in drug discovery. *Med. Res. Rev.* 20, 197–202
17 DeRisi, J. *et al.* (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14, 457–460
18 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537
19 Alizadeh, A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511
20 Alon, U. *et al.* (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6745–6750
21 Hedenfalk, I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539–548
22 Ono, K. *et al.* (2000) Identification by cDNA microarray of genes involved in ovarian carcinogenesis. *Cancer Res.* 60, 5007–5011
23 Hecht, S.S. (1999) Tobacco smoke carcinogens and lung cancer. *J. Natl. Cancer Inst.* 91, 1194–1210
24 Keane, J. *et al.* (2000) Virulent *Mycobacterium tuberculosis* strains evade apoptosis of infected alveolar macrophages. *J. Immunol.* 164, 2016–2020
25 McMaster, G. (2000) The impact of Genomics' technologies on pharmaceutical research. *Med. Res. Rev.* 20, 187–188
26 Harris, T. (2000) Genetics, genomics, and drug discovery. *Med. Res. Rev.* 20, 203–211
27 Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846
28 Zhu, H. and Snyder, M. (2001) Protein arrays and microarrays. *Curr. Opin. Chem. Biol.* 5, 40–45
29 Fung, E.T. *et al.* (2001) Protein biochips for differential profiling. *Curr. Opin. Biotechnol.* 12, 65–69
30 Parekh, R.B. and Rohlff, C. (1997) Post-translational modification of proteins and the discovery of new medicine. *Curr. Opin. Biotechnol.* 8, 718–723
31 Mann, M. and Pandey, A. (2001) Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* 26, 54–61
32 Yates, J.R., III (1998) Database searching using mass spectrometry data. *Electrophoresis* 19, 893–900
33 Rout, M.P. *et al.* (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* 148, 635–651
34 Peltier, J.B. *et al.* (2000) Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* 12, 319–341
35 Neubauer, G. *et al.* (1997) Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 94, 385–390
36 Rain, J.C. *et al.* (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–215
37 Walhout, A.J. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122
38 Ito, T. *et al.* (2000) Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci. U. S. A.* 97, 1143–1147
39 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
40 Sche, P.P. *et al.* (1999) Display cloning: functional identification of natural product receptors using cDNA-phage display. *Chem. Biol.* 6, 707–716
41 Cochrane, D. *et al.* (2000) Identification of natural ligands for SH2 domains from a phage display cDNA library. *J. Mol. Biol.* 297, 89–97

42 Fadok, V.A. *et al.* (2000) A receptor for phosphatidylserine-specific clearance of apoptotic cells. *Nature* 405, 85–90
43 Rader, C. and Barbas, C.F. (1997) Phage display of combinatorial antibody libraries. *Curr. Opin. Biotechnol.* 8, 503–508
44 Sidhu, S.S. (2000) Phage display in pharmaceutical biotechnology. *Curr. Opin. Biotechnol.* 11, 610–616
45 Hanes, J. *et al.* (2000) Selecting and evolving functional proteins *in vitro* by ribosome display. *Methods Enzymol.* 328, 404–430
46 Roberts, R.W. and Szostak, J.W. (1997) RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* 94, 12297–12302
47 Lu, Z. *et al.* (1995) Expression of thioredoxin random peptide libraries on the *Escherichia coli* cell surface as functional fusions to flagellin: a system designed for exploring protein-protein interactions. *Biotechnology.* (N.Y.) 13, 366–372
48 Boder, E.T. and Wittrup, K.D. (2000) Yeast surface display for directed evolution of protein expression, affinity, and stability. *Methods Enzymol.* 328, 430–444
49 Norman, T.C. *et al.* (1999) Genetic selection of peptide inhibitors of biological pathways. *Science* 285, 591–595
50 Li, M. (2000) Applications of display technology in protein analysis. *Nat. Biotechnol.* 18, 1251–1256
51 Williams, C. (2000) Biotechnology match making: screening orphan ligands and receptors. *Curr. Opin. Biotechnol.* 11, 42–46
52 Colas, P. (2000) Combinatorial protein reagents to manipulate protein function. *Curr. Opin. Chem. Biol.* 4, 54–59
53 Gaytan, P. *et al.* (1998) Combination of DMT-mononucleotide and Fmoc-trinucleotide phosphoramidites in oligonucleotide synthesis affords an automatable codon-level mutagenesis method. *Chem. Biol.* 5, 519–527
54 Cho, G. *et al.* (2000) Constructing high complexity synthetic libraries of long ORFs using *in vitro* selection. *J. Mol. Biol.* 297, 309–319
55 Weng, Z. and DeLisi, C. (2000) Amino acid substitutions: effects on protein stability. In *Encyclopedia of Sciences*
56 Kieke, M.C. *et al.* (1999) Selection of functional T cell receptor mutants from a yeast surface-display library. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5651–5656
57 Stemmer, W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370, 389–391
58 Forrest, S. (1993) Genetic algorithms – principles of natural selection applied to computation. *Science* 261, 872–878
59 Chang, C.C. *et al.* (1999) Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* 17, 793–797
60 Zeng, J. (2000) Mini-review: computational structure-based design of inhibitors that target protein surfaces. *Comb. Chem. High Throughput Screen.* 3, 355–362
61 Gane, P.J. and Dean, P.M. (2000) Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* 10, 401–404
62 Gschwend, D.A. *et al.* (1997) Specificity in structure-based drug design: identification of a novel, selective inhibitor of *Pneumocystis carinii* dihydrofolate reductase. *Proteins* 29, 59–67
63 Cho, M.J. and Juliano, R. (1996) Macromolecular versus small-molecule therapeutics: drug discovery, development and clinical considerations. *Trends Biotechnol.* 14, 153–158
64 Janin, J. (1996) Quantifying biological specificity: the statistical mechanics of molecular recognition. *Proteins* 25, 438–445
65 Sternberg, M.J. *et al.* (2000) Protein–protein docking. Generation and filtering of complexes. *Methods Mol. Biol.* 143, 399–415
66 Jackson, R.M. and Sternberg, M.J.E. (1995) A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol.* 250, 258–275
67 Vakser, I.A. and Aflalo, C. (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. *Proteins* 20, 320–329
68 Vajda, S. *et al.* (1994) Effect of conformational flexibility and solvation on receptor–ligand binding free energies. *Biochemistry* 33, 13977–13988
69 Weng, Z. *et al.* (1996) Prediction of protein complexes using empirical free energy functions. *Protein Sci.* 5, 614–626
70 Weng, Z. *et al.* (1998) Computational determination of the structure of rat Fc bound to the neonatal Fc receptor. *J. Mol. Biol.* 282, 217–225
71 Zhang, C. *et al.* (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* 267, 707–726
72 Zhang, C. *et al.* (1997) Consistency in structural energetics of protein folding and peptide recognition. *Protein Sci.* 6, 1057–1064
73 Gabb, H.A. *et al.* (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* 272, 106–120
74 DeLano, W.L. *et al.* (2000) Convergent solutions to binding at a protein-protein interface. *Science* 287, 1279–1283
75 Laskowski, R.A. *et al.* (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
76 Brady, G.P. Jr and Stouten, P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* 14, 383–401
77 Gallet, X. *et al.* (2000) A fast method to predict protein interaction sites from sequences. *J. Mol. Biol.* 302, 917–926
78 Ringe, D. (1995) What makes a binding site a binding site? *Curr. Opin. Struct. Biol.* 5, 825–829
79 Farber, G.K. (1999) New approaches to rational drug design. *Pharmacol. Ther.* 84, 327–332
80 Katchalski-Katzir, E. *et al.* (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U. S. A.* 89, 2195–2199
81 Palma, P.N. *et al.* (2000) BiGGER: A new (soft) docking algorithm for predicting protein interactions. *Proteins* 39, 372–384
82 Ritchie, D.W. and Kemp, G.J. (2000) Protein docking using spherical polar Fourier correlations. *Proteins* 39, 178–194
83 Gardiner, E.J. *et al.* (2001) Protein docking using a genetic algorithm. *Proteins* 44, 44–56
84 Vakser, I.A. *et al.* (1999) A systematic study of low-resolution recognition in protein–protein complexes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 8477–8482
85 Strynadka, N.C. *et al.* (1996) Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. *Nat. Struct. Biol.* 3, 233–239
86 Dixon, J.S. (1997) Evaluation of the CASP2 docking section. *Proteins* (Suppl.) 1, 198–204
87 Jackson, R.M. *et al.* (1998) Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.* 276, 265–285
88 Camacho, C.J. *et al.* (2000) Scoring docked conformations generated by rigid-body protein-protein docking. *Proteins* 40, 525–537
89 Burley, S.K. *et al.* (1999) Structural genomics: beyond the human genome project. *Nat. Genet.* 23, 151–157
90 Hellinga, H.W. (1998) Computational protein engineering. *Nat. Struct. Biol.* 5, 525–527
91 DeGrado, W.F. *et al.* (1999) De novo design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* 68, 779–819
92 Gordon, D.B. *et al.* (1999) Energy functions for protein design. *Curr. Opin. Struct. Biol.* 9, 509–513
93 Dahiyat, B.I. and Mayo, S.L. (1997) De novo protein design: fully automated sequence selection. *Science* 278, 82–87
94 Malakauskas, S.M. and Mayo, S.L. (1998) Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.* 5, 470–475
95 De Maeyer, M. *et al.* (1997) All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* 2, 53–66
96 Marshall, S.A. and Mayo, S.L. (2001) Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.* 305, 619–631
97 Ross, S.A. *et al.* (2001) Designed protein G core variants fold to native-like structures: sequence selection by ORBIT tolerates variation in backbone specification. *Protein Sci.* 10, 450–454
98 Voigt, C.A. *et al.* (2000) Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* 299, 789–803
99 Wernisch, L. *et al.* (2000) Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.* 301, 713–736
100 Looger, L.L. and Hellinga, H.W. (2001) Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J. Mol. Biol.* 307, 429–445
101 Davis, M.M. *et al.* (1998) Ligand recognition by alpha beta T cell receptors. *Annu. Rev. Immunol.* 16, 523–544
102 Weng, Z. and DeLisi, C. (1998) Toward a predictive understanding of molecular recognition. *Immunol. Rev.* 163, 251–266
103 Yang, W.P. *et al.* (1995) CDR walking mutagenesis for the affinity maturation of a potent human anti-HIV-1 antibody into the picomolar range. *J. Mol. Biol.* 254, 392–403
104 Hanes, J. *et al.* (2000) Picomolar affinity antibodies from a fully synthetic naive library selected and evolved by ribosome display. *Nat. Biotechnol.* 18, 1287–1292