suggested that the advantages of the system are high expression levels, linear scale-up and the low cost of goods. Additionally, the protein is produced in a purification-friendly environment and the post-translational modifications are human-like. PPL has used the technology to produce a number of proteins, one of which (α1 antitrypsin) is in Phase III clinical trials.

Concluding remarks

The take-home message for the industry outsider was that the lessons learned from the mistakes made during the initial attempts to bring protein therapeutics to the market are now being addressed at a very early stage in the discovery process. Parallels can be drawn with the experience gained by the small-molecule therapeutic discovery process in that it is now recognized that multidimensional optimization of the compounds throughout the discovery process will lead to decreased costs and higher success rates in the clinic. The industry as a whole has laid the foundations for financial success but, most importantly, it is poised to provide many new medicines to treat the various medical needs.

Biocomputing: impact of the genomic revolution

Rebecca Lawrence, News & Features Editor

The annual meeting of the *Pacific Symposium on Biocomputing*, held in Mauna Lani (HI, USA) was greatly oversubscribed this year due to the growing importance of computational biology and bioinformatics in the post-genomic era. The meeting covered a wide range of aspects, from DNA and protein structure, protein–DNA interactions and expression, protein evolution, and human genome variation to phylogenetics, high-performance computing, natural language processing and bioethics. This report will provide an overview of the main talks relating to drug discovery. The full papers presented at the meeting are available in the symposium book[1].

The keynote speaker, David Haussler, provided an overview of the effort to produce the working draft of the human genome sequence. He admitted there will be some cross-contamination present in the sequence as all the data that passes the NCBI filters is used in the final assembly. He also pointed out that the bioinformatics tools used are not perfect and miss-alignments have created some artefactual duplication. He reiterated the plan to complete the sequence by 2003, and that the next steps will be to identify the complete set of human genes, explore gene regulation, research human, mammalian and vertebrate diversity, and connect genomic data to clinical data.

Comparing sequences

A range of different computational approaches for sequence comparison was discussed. William Martins (Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA) and colleagues have developed a dynamic programming algorithm that uses a rigorous mathematical approach based on a multi-threading system. This system can be scaled up and actually performs better with longer sequences, providing the capability to compare whole genomes. A further advantage of this method is that it enables the visual display of the matching sections of the genomes.

A new algorithm based on Monte Carlo multiple sequence alignment techniques was used by C. Guda (San Diego Supercomputer Center, University of California San Diego, La Jolla, CA, USA) and colleagues. Four different types of moves were designed to generate random changes in the alignment – shifting (shifts residues in one randomly chosen structure); expanding (expands the alignment block by acquiring one residue); shrinking (shrinks the alignment block by removing one residue); and splitting-and-shrinking (splits longer blocks in two and then shrinks one of the new blocks). Changes in distance-based scores were examined for each trial move and changes were only accepted if the alignment score increased. This process was repeated until the alignment scores converged. Each step only examines changes of one residue at a time, but many were concerned that more global changes might be missed that could have significantly improved the score.

The use of Kestrel single-instruction multiple-data (SIMD) parallel processor methods for sequence analysis was

1359-6446/01/$ – see front matter PII: S1359-6446(01)01737-8

discussed by Richard Hughey (Department of Computer Engineering, University of California Santa Cruz, CA, USA). Three-point pharmacophore fingerprinting, which examines the distances between each possible pair of atoms (or atom groups), produces trillions of atom triplet calculations for large libraries and is therefore impractical to carry out on a serial processor. Hence, Hughey's team have developed a parallel processor method by simultaneously processing multiple conformations for molecular fingerprinting. This method was then tested on data produced by Affymetrix, where they found it was 35-fold faster than standard molecular fingerprinting. The group are now trying to further optimize this system to enable storage of all the information on a single chip.

Gene expression

A method using microarray techniques to analyse whole genome gene expression was discussed by Xiaole Liu (Stanford Medical Informatics, Stanford University, Stanford, CA, USA). Liu and colleagues used a Gibbs sampling technique (BioProspector), which isolates one sequence from a motif, scores each possible segment of the sequence, samples the segment and then updates the motif. This process is repeated until the whole motif is produced.

The original Gibbs sampler method relies on each sequence containing only one sample copy of the motif. Because some input sequences contain no copies of the motif while others contain many copies, the group has used two thresholds in the sampler. Everything that is above the high threshold will automatically be added to the alignment. The low threshold starts at zero and is gradually increased. Only the highest scoring segment between the two thresholds will be added to the final alignment.

To understand the mechanism of a class of genes, it is important to keep track of not only what motifs occur but also how often they occur, their order and spacing, and the promoter classification. With this in mind, Bill Grundy (Department of Computer Science, Columbia Genome Center, Columbia University, New York, NY, USA) and colleagues have devised a method of supervised learning for promoter region-based classification of genes. This involves discovering conserved motifs in training set sequences, building a motif-based hidden Markov model, computing the gradients of the model parameters with respect to the training set sequences, and then training the model. Using this method for two classes of genes from the budding yeast, *Saccharomyces Cerevisiae*, has enabled them to correctly classify the promoters involved.

Protein structure and evolution

This session was introduced by Richard Goldstein (Chemistry Department and Biophysics Research Division, University of Michigan, Ann Arbor, MI, USA) who highlighted that one of the key post-genomic steps will be to discover the properties of the proteins and how these are determined by amino acid properties. One method is to change one amino acid at a time and examine how the protein structure changes but this method cannot be used to learn the basis of the changes. The alternative is to examine variation and selection between individuals and then work back to elucidate the selective pressures acting on these proteins.

Goldstein's team have developed a substitution model that considers the amino acid residues observed in each position and constructs a separate substitution matrix for every location. As the 'rules' that these locations follow is unknown, sites have to be assigned probabilistically. The resulting Hidden States Model can identify which locations in a protein have similar substitution rates. The main drawback of this approach is the large number of adjustable parameters it produces that have to be simultaneously determined.

Their studies found that fast-changing sites had hydrophobic pressure outside but hydrophilic pressure inside while slow-changing sites were the reverse. Furthermore, for all secondary structures, hydrophobicity was the dominant factor for exposed residues while the bulk index was dominant for buried residues. The strongest contributions of hydrophobicity were from exposed α-helices, followed by exposed β-sheets, and then turns and coils. It was suggested that it would be interesting to also look at a group of residues rather than just one residue at a time.

Keith Dunker (School of Molecular Biosciences, Washington State University, Pullman, WA, USA) described the compilation of three primary and one derivative database of intrinsically disordered proteins to investigate the determinants of protein order and disorder. The amino acid compositions from the combined four disordered databases were compared with those from a database of ordered proteins. The disordered databases were found to be significantly depleted of 8 amino acids and significantly enriched of a further 8 amino acids, suggesting that there is a set of order-promoting and disorder-promoting amino acids.

The group also identified the top ten attributes of the amino acids that correlated as little as possible with each other and that promoted order or disorder. Of these ten properties, the most important property is the 14 Å contact number. Of the other nine properties, one is the coordination number, four are associated with hydrophobicity, one is associated with polarity and three are related to the propensity of the amino acids to form β-strands. The team is now working on determining whether there are any relationships between protein function and types of protein disorder.

Linking genotypes to clinical phenotypes

Francisco de la Vega (Applied Biosystems, Foster City, CA, USA) gave an overview

of this field and the technologies and data required. He suggested future challenges include technological reproducibility and quality control at high-throughput, the accurate analysis of data, and good data integration. Peter Park (Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA) then discussed the use of microarray data for predicting clinical phenotypes and the difficulties in knowing which genes to use and the unusual statistical problems of there being more variables than subjects. His group has therefore developed a nonparametric scoring algorithm, which uses ranks rather than actual data and is less sensitive to measurement errors. Through scoring of each gene based on samples of known classes, they discussed how a small set of informative genes can be isolated from the large amount of gene expression level data produced by the large quantity of genes.

Another problem that often arises is the difficulty in determining what counts as a significant difference in gene expression analyses. Atul Butte (Children's Hospital Informatics Program, Boston, MA, USA) commented on the arbitrary threshold differences used by researchers and the problems of comparing data between microarray chips. Butte's group tested this by comparing RNA data from four patients with glucose intolerance between two sets of Affymetrix Hu35K microarrays. Initially, they found seemingly high reproducibility between the two sets of chips with correlation coefficients of intrapatient-repeated measurements being 0.76–0.84. However, they found poor reproducibility when they examined interpatient-fold differences, ranging from 0.01 to 0.09. In one instance, they found that a gene was expressed tenfold higher or tenfold lower in patient 1 compared with patient 3 depending on which set of chips were being used. His explanation for this was that the measurements are not 'stable': noise changes expression measurements as the numbers are so small. He suggested this could be overcome by removing 'noisy' patients and by using a smaller range of insignificance. However, he concluded that the best strategy would be different for a small group working in academia where money is a significant factor, compared with a large pharmaceutical company where there are sufficient funds to follow up false-positives.

Pharmacogenomic database

Russ Altman (Stanford Medical Informatics, Stanford University, Stanford, CA, USA) provided an overview of the pharmacogenomic database (PharmGKB; http://www.pharmgkb.org/) being set-up in parallel with a private pharmacogenomic effort to provide a publicly available repository of pharmacogenomic information. The database will be run by Teri Klein (Standford Medical Informatics) and, when complete, will include genomic information, alleles, clinical phenotypes, molecular and cellular phenotypes, molecules, drug response systems, drugs and impact of the environment. Groups that do not have enough bioinformatics can submit the gene sequence information and the database group will then do the necessary bioinformatics for them. The groups currently involved in the project are working on asthma, cancer, transporter molecules, oestrogen-related genes, depression and Phase II metabolism enzymes.

Bioethics

An interesting discussion session was held on bioethics. Pierre Baldi (Department of Information and Computer Science and Department of Biological Chemistry, University of California Irvine, CA, USA) overviewed some of the recent developments in computers and in science. These included the development of computers that will soon exceed the computational storage power of the human brain; connectivity that approaches science fiction; and genetically engineered animals such as tigrons and ligras (part lion, part tiger) and glowing mice and plants. He suggested that the challenges for the future would be to identify what it actually means to be human and the blurring of the silicon/carbon boundary.

Larry Hunter (University of Colorado, Boulder, CO, USA) highlighted the problems of deciphering 'good' from 'bad' genes. Obvious 'good' genes might be for disease resistance, physical health or mental ability, while 'bad' genes might be for cancer and heart disease, mental and physical diseases and disabilities. However, the boundaries are more complicated with, for example, the *globin* gene, as when heterozygous, the gene variant protects against malaria, but when homozygous, it produces sickle cell anaemia. Other determinants include how problematic some 'bad' genes really are. For example, poor eyesight might take 20 years to develop by which time a non-genetic intervention might be available that means genetic intervention is not required.

However, James Sikela (University of Colorado Health Sciences Center, Boulder, CO, USA) pointed out that all courses of action have consequences, including doing nothing at all about the genetic revolution. The challenges are to anticipate the consequences, build consensus on goals and devise effective strategies. He concluded the session by reminding everyone that we all have a responsibility to carefully weigh up all sides of our actions relating to the use of genetic information before we act.

Acknowledgements

I would like to thank Keith Dunker for his valuable comments on this article.

Reference

1 Altman, R.B. *et al.* (eds) (2001) Pacific Symposium on Biocomputing 2001, World Scientific Publishing Co.