Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

SECTION 5

GENETICS/GENOMICS INTERFACES

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 15

Gene Expression Informatics and Analysis

ANTOINE H. C. VAN KAMPEN*, JAN M. RUIJTER, BARBERA D. C. VAN SCHAIK, HUIB N. CARON, and ROGIER VERSTEEG

Academic Medical Center, University of Amsterdam
Meibergdreef 9, 1105 AZ Amsterdam
The Netherlands

* corresponding author

15.1 INTRODUCTION

Unravelling the molecular mechanisms in living cells is one of the major challenges in current biology. Understanding these mechanisms will help us to recognize and finally treat a range of diseases such as cancer. Gene expression profiling provides one approach to study cellular processes at the gene level. There are many approaches for the measurement of gene expression, at a single gene level, technologies such as RT-PCR and Taqman provide detailed gene expression profiles across a defined range of tissues (Heid *et al*., 1996; Riedy, *et al*., 1995). At a genome-wide level, Serial Analysis of Gene Expression (SAGE) (Velculescu *et al*., 1995) and DNA microarrays (Lockhart *et al*., 1996; Schena *et al*., 1995; Zammatteo *et al*., 2000) enable the simultaneous measurement of the expression of thousands of genes in a single tissue.

The advances in physical transcript mapping afforded by the recently released draft sequence of the human genome (Lander *et al*., 2001; Venter *et al*., 2001) creates a new opportunity to combine genome-wide gene profiling and gene mapping efforts. Combination of gene expression data with gene positions will further unravel molecular mechanism in the cell. This chapter focuses on *in silico* sources of gene expression data, such as SAGE and the Human Transcriptome Map (HTM; Caron *et al*., 2001), for the evaluation of gene expression across loci, specifically addressing the needs of positional cloning and cancer genetics.

Cancer results from changes in DNA sequence, which are reflected in altered amino acid sequences of the corresponding proteins or changes in protein expression levels, either of which ultimately changes cell function (King, 2000). These DNA changes can include (relatively) small mutations involving substitutions, insertions or deletions of bases, but also gross changes in DNA content per nucleus manifested as chromosome rearrangements or as gene amplifications.

DNA changes that eventually lead to cancer may be reflected in the expression levels of the corresponding genes or in the expression levels of genes that are directly or indirectly regulated by the mutated gene(s). Consequently, comparison and analysis of gene expression profiles of normal and tumour tissue at different stages of carcinogenesis helps to increase our knowledge of the molecular biology of cancer. Although the translation of expression profiles to relevant biological information is still one of the major challenges in biology and bioinformatics, integral gene expression analysis is already used extensively in cancer research (e.g. Alizadeh *et al*., 2000; Ben-Dor *et al*., 2000; Cole *et al*., 1999; Golub *et al*., 1999; Hastie *et al*., 2000; Spieker *et al*., 2001; Yeang *et al*., 2001; Zhang *et al*., 1997).

One problem that occurs in the comparison of gene expression profiles for normal and tumour tissues is the large number of genes that are differentially expressed. Not all these genes are interesting candidates for further investigation, most are not directly implicated in carcinogenesis, but instead they may be part of the multiple downstream pathways which are activated during carcinogenesis, including for example, responses of the cell to stress or apoptosis. Therefore, to facilitate the identification of candidate genes it could be useful to select those genes that are positioned at aberrant regions of chromosomes. Many such regions are already known for different types of cancer or they can easily be detected

by screening tumour material (Mitelman *et al*., 2001). For example, the embryonal tumour neuroblastoma shows common genetic aberrations such as the amplification of the MYCN oncogene (Schwab *et al*., 1983) and loss of chromosome 1p (Brodeur *et al*., 1977).

The Human Transcriptome Map (HTM) was specifically developed to enable the comparison of expression levels of genes in such regions. The HTM provides a clear example of a project that was initiated with a simple question: 'Is it possible to develop a tool that guides the identification of candidate genes from chromosomal regions known to be involved in neuroblastoma (or other cancers) from genome-wide gene expression profiles obtained with Serial Analysis of Gene Expression (SAGE)?'. To answer this question, the HTM integrates the position of human genes on chromosomes with genome-wide expression profiles provided by SAGE (Velculescu *et al*., 1995).

Although the HTM seems a straightforward integration of a gene mapping database and SAGE expression profiles, the development of such an application is actually quite complex as will be shown in this chapter. Numerous aspects had to be considered during the development of the HTM such as the development of sequence analysis algorithms as part of the SAGE analysis, the application of statistical methods to analyse the data and the development of a relational database to enable the integration of data from different (public) resources.

HTM was initially developed for the selection of candidate genes but it also provides more fundamental insight into the organization of the human genome. Inspection of the expression profiles for all chromosomes reveals an intriguing pattern of domains of genes with an above-average expression in each tissue. These domains were named RIDGEs (Regions of Increased Gene Expression) and understanding them may further advance our knowledge of normal organization of the genome and of cancer.

In using computational tools such as HTM it is important to have a basic understanding of the underlying principles of the tool to avoid misinterpretation of results. In general, software applications are used as 'black boxes' that give answers to questions when data is put in. On the other hand, once underlying technologies and principles are understood, it is possible to identify new possibilities for application of tools or generation of ideas for the development of new tools. The use of public biological databases also requires caution since they may contain errors, ambiguous data or data may be missing (e.g. Karp, 1998; Karp *et al*., 2001). Understanding the nature of the data contained in these databases will facilitate the interpretation of the results obtained. This chapter provides some examples of the issues and pitfalls that are involved in the construction and the application of gene expression analysis tools.

This chapter focuses specifically on issues related to the Human Transcriptome Map and therefore it necessarily concentrates on SAGE technology and bioinformatics of SAGE analysis. However, it will become clear that the overall approach, technologies and problems described during the development of the HTM are not specific for SAGE analysis but also apply to technologies such as DNA microarrays. In this chapter we will describe the SAGE and DNA microarray technologies and discuss some important differences between these two technologies. We will introduce the Cancer Genome Anatomy Project, which has made several tools and databases for gene expression analysis available via the internet. We will discuss the processing and statistical analysis of SAGE data and explain the data integration process that was required to construct the HTM.

15.2 TECHNOLOGIES FOR THE MEASUREMENT OF GENE EXPRESSION

A range of methods are available to measure gene expression or changes in gene expression. In this section the SAGE and DNA microarray technologies are described since these are the primary methods used for genome-wide profiling.

15.2.1 Serial Analysis of Gene Expression (SAGE)

Serial Analysis of Gene Expression (SAGE; Velculescu *et al*., 1995) is a technique used to construct quantitative genome-wide gene expression profiles (van Limpt *et al*., 2000; Porter *et al*., 2001; Scott and Chrast, 2001; Velculescu *et al*., 2000). Three principles underlie the SAGE methodology (Figure 15.1):

(1) A short 10-base pair sequence tag contains sufficient information to uniquely identify a transcript provided that this tag is obtained from a unique position within each transcript (there are many more possible tags (4^{10} = 1,048,576) than human genes).
(2) Sequence tags can be linked together to form long serial molecules (concatemers) that can be cloned and sequenced.
(3) Counting of the number of times a particular tag is observed provides the expression level of the corresponding transcript.

Figure 15.1 Serial Analysis of Gene Expression (SAGE; Velculescu *et al*., 1995). mRNA is extracted from a cell tissue sample. Subsequently, a 10-base pair tag that is right to the most 3′ CATG site is extracted from each transcript by using the NlaIII restriction enzyme. These tags are then ligated to ditags, which are amplified and linked to form concatemers containing approximately 30 tags. These concatemers are then cloned and sequenced.

The sequenced concatemers consist of ditags and include approximately 30 to 40 tags. The sequenced concatemers are the starting point for data processing, which is explained in more detail in Section 15.4. The number of tags that is obtained in a SAGE experiment ranges from 10,000 to over 100,000. The frequency of a tag directly reflects the fraction of the corresponding transcript in the cell. In other words, if a particular tag is observed 25 times in a SAGE library that consists of 50,000 tags, then the number of corresponding transcripts is also 25 per 50,000 transcripts in the cell. In this sense, SAGE provides an 'absolute' expression level.

15.2.2 DNA Microarrays

The principle of a DNA microarray experiment is to hybridize labelled cDNA to DNA sequences that are immobilized on a solid surface in an ordered array. The labelled cDNA is often referred to as the target and the immobilized DNA sequences as the probe. A DNA microarray allows the detection and quantification of thousands of transcripts simultaneously. Two main types of DNA microarrays can be distinguished according to the arrayed material. The first type is the cDNA microarray in which the probes are usually products of the polymerase chain reaction generated from cDNA libraries or clone collections (Bowtell, 1999; Brown and Botstein, 1999; Schena *et al*., 1995, 1996). These probes are spotted onto glass slides or nylon membranes at defined positions. The second type or arrays are the oligonucleotide arrays for which short 20–25mers are synthesized *in situ* by photolithography onto silicon wafers (GeneChip™ technology of Affymetrix (Lockhart *et al*., 1996)). Alternatively, pre-synthesized oligonucleotides can be printed onto glass slides (Okamoto *et al*., 2000, Zammatteo *et al*., 2000). For target preparation, mRNA from cells or tissue is extracted, which is converted to cDNA and labelled. The target is then hybridized to the DNA probes on the array and detected by phospho-imaging or fluorescence scanning. In the case of fluorescence, two fluorescent dyes with different colours (Cy3 and Cy5) are used to label the cDNAs from two different cell populations. The resulting two targets are mixed and hybridized to the same array, which results in competitive binding of the target to the spotted probe sequences. Subsequently, the array is scanned using two different wavelengths, corresponding to the two dyes and the intensity of each spot in both channels is 'mixed' *in silico*. This results in an expression level, relative to the chosen control condition, for each gene that is represented on the array (see Chapter 9 Section 9.6, for some examples of output from oligonucleotide arrays).

15.2.3 Comparison of SAGE and DNA Microarrays

The SAGE and DNA microarray technologies differ in several important ways. SAGE measures expression levels that directly reflect the fraction of mRNAs in the cell, i.e. SAGE produces 'absolute' expression levels. In contrast, the DNA microarray technique measures expression levels relative to a control condition. Consequently, different SAGE libraries can be directly compared because the expression levels do not depend on the use of a reference mRNA or experimental conditions, while DNA microarray experiments can only be compared if they have been measured relative to the same control tissue under the same conditions. For the same reason, the gene expression levels within one SAGE library can be directly compared, while expression levels obtained for genes on one DNA microarray cannot be compared due to differences in labelling and hybridization efficiency of individual genes. Another difference between SAGE and DNA microarrays comprises

the genes that can be measured in an experiment. In a DNA microarray experiment one only measures the genes for which the array contains probes, while SAGE in principle measures every mRNA in the sample. Consequently, SAGE is very suitable for discovering new genes, although low abundance transcripts are only likely to appear in large SAGE libraries. The DNA microarray is, however, very suitable for quickly screening cells or tissues for the expression of a pre-selected set of genes. A disadvantage of SAGE is that the extracted mRNA tags need to be identified *in silico* (Section 15.4) while for DNA microarrays it is already known which probes (genes) are on the array. Furthermore, the construction of a SAGE library requires much more effort than carrying out a DNA microarray experiment once the array has been printed.

15.3 THE CANCER GENOME ANATOMY PROJECT (CGAP)

The Cancer Genome Anatomy Project (CGAP; Lal *et al*., 1999; Lash *et al*., 2000; Riggins and Strausberg, 2001; Schaefer *et al*., 2001; Strausberg *et al*., 1997, 2000) is a project of the National Cancer Institute (NCI). Their main objective is to decipher the molecular mechanism of cancer. For this goal, information is gathered from different resources such as gene expression data, aberrations of chromosomes, gene variation and biochemical pathways. CGAP collaborates with the National Centre for Biotechnology Information (NCBI) to develop computational technologies for the management and analysis of these large amounts of data. All data and programs for analysis are made available via the internet (cgap.nci.nih.gov).

The HTM makes extensive use of two CGAP resources. Firstly, HTM includes the SAGE libraries that were constructed as part of the CGAP project. Secondly, HTM algorithms use the SAGEmap tag-to-gene mapping as a starting point for constructing an improved tag-to-gene mapping (Section 15.4.2). These mappings are used for SAGE tag identification. In addition, several other CGAP tools and databases are regularly used during the analysis of SAGE data. Therefore, this section provides a brief overview of the resources offered by CGAP.

The CGAP resource contains cDNA and SAGE libraries of normal cells and cancer cells in different stages. These libraries include the 3′ and 5′ clones of cDNAs from the dbEST database (Boguski *et al*., 1993), the CGAP subset of dbEST, the Mammalian Gene Collection (MGC) subset of dbEST, randomly cloned cDNAs from the ORESTES (Open Reading Frame EST sequencing) project and SAGE libraries. The MGC is an NIH initiative that supports the production of cDNA libraries, clones and sequences (Strausberg *et al*., 1999). The goal of the MGC is to provide a complete set of full-length (open reading frame) sequences and cDNA clones of expressed genes for human and mouse. The ORESTES project aims for the completion of gene annotation by sequencing randomly primed cDNAs (Pandey, 2001). CGAP also supports the generation of SAGE libraries and their sequencing to obtain gene expression profiles of normal, pre-cancer, and cancer cells, which resulted in high-quality SAGE gene expression profiles for a range of normal and tumour tissues. The generation of these profiles still continues. At present over 140 SAGE libraries are available from the SAGEmap database (Lal *et al*., 1999; Lash *et al*., 2000) including more than 5 million tags.

The CGAP Library Finder Tool retrieves any cDNA library from dbEST or SAGE libraries. The search can be narrowed to the CGAP, MGC or ORESTES subsets. A query first returns a single library or a list of libraries, each of which is linked to its own Library Info page where details of the library and its preparation can be found. The Library Finder

Tool allows the retrieval of libraries according to tissue type, tissue preparation, tissue histology, library protocol and library name.

CGAP offers a range of tools to examine gene expression data from their cDNA or SAGE collection. The Gene Library Summarizer (GLS) generates unique and non-unique genes expressed in a single cDNA library or library group. It then identifies the genes in each of these groups as known or unknown. The cDNA xProfiler is a tool that compares gene expression between two pools of libraries by counting the number of clones in the library. The Digital Gene Expression Displayer (DGED) is a tool that compares gene expression between two pools of libraries. In contrast to the cDNA xProfiler that counts clones, the DGED treats the presence of a gene in a library pool as a matter of degree. It compares the 'degree' of presence of a gene in pool A with its 'degree' of presence in pool B by using a chi-squared test. The SAGEmap xProfiler performs differential-type analyses on (pooled) SAGE libraries. Similar libraries can be placed into one of two groups based on their characteristics (e.g. normal colon and colon cancer). Comparisons are then made between the two groups using a statistical test developed specifically for SAGE data (Lash *et al*., 2000). The SAGEmap Virtual Northern (vNorthern) tool has been designed to accept mRNA or EST sequences as input. Possible tags are then extracted from this sequence and links provided to access the data from the various SAGE libraries currently represented on the SAGEmap website.

CGAP also provides access to the Mitelman database of chromosome aberrations in cancer (Mitelman *et al*., 2001). This database contains manually selected data from about 40,000 scientific articles and is organized as three distinct sub-databases. The sub-database 'Cases' contains the data that relates chromosomal aberrations to specific tumour characteristics in individual patient cases. The sub-database of 'Molecular Biology and Clinical Associations' contains no data from individual patient cases. Instead, the data is pulled from studies with distinct information about molecular biology or clinical associations. The molecular biology associations relate chromosomal aberrations and tumour histologies to genomic sequence data, while clinical associations relate chromosomal aberrations and tumour histologies to clinical variables such as prognosis, tumour grade and patient characteristics. The 'Reference' sub-database contains all the references culled from the literature.

Another tool to examine the chromosomes uses the CGAP FISH-mapped BACs, which are BAC clones that are mapped both cytogenetically by FISH and physically by STSs to the human genome. Genetic and physical SNP maps are available, which show the genetic and physical locations of confirmed, validated and predicted SNPs per individual chromosome.

CGAP also includes the graphical biochemical pathway maps from KEGG (Kanehisa and Goto, 2000) and BioCarta (www.biocarta.com). The entities on these maps are linked to the above-mentioned CGAP resources.

15.4 PROCESSING OF SAGE DATA

The processing of SAGE data generally consists of three steps. First a list of tags is compiled from the concatemer sequences. Secondly, the SAGE tags are identified and finally the expression levels can be compared statistically. The extraction of tags from the concatemer sequences is straightforward since each concatemer consists of ditags that are separated by the CATG sequence. Each ditag contains one tag in the $5' \rightarrow 3'$ (sense) direction and a second tag in the $3' \rightarrow 5'$ (complementary-reverse) orientation. The ditags

are extracted from the concatemers and duplicate ditags are removed because they are most likely experimental artifacts (Velculescu *et al*., 1995). The length of the resulting ditags must be between 20 and 24 bp. Shorter and longer ditags are discarded as experimental artefacts. Subsequently, from each extracted ditag the sense and complementary-reverse (which is converted to a sense tag) tags are extracted and added to the list of SAGE tags. The number of times that a tag occurs in this list directly reflects the expression level of the corresponding transcript.

As a result of this experimental procedure the association between tag and transcript from which the tag is extracted is lost. Consequently, after compiling the tag list (i.e. gene expression profile) each tag in this list has to be identified by matching it against a tag-to-gene map. This tag-to-gene mapping database must first be compiled by electronically extracting a tag from each mRNA/EST sequence in the GenBank database and subsequently storing the annotated tag in the tag-to-gene mapping. The compilation of this tag-to-gene mapping is one of the crucial steps in the SAGE analysis.

The CGAP SAGEmap tag-to-gene mapping (Lal *et al*., 1999) is an example of such mapping. Typical entries in this tag-to-gene mapping look something like the following:

AAAAATACAA 5/EST/+3_label *43744 ESTs* AI093649, AI263776, N26090, N67808 (4 6)
TATTAGGATA 5/EST/+3_label *43744 ESTs* AI434789, AI813305, AW271602 (3 3)
AAAAAATACA 1/mRNA/+orient *1119 nuclear receptor subfamily 4, group A* D85245 (1 1)
AAAAAATACA 2/EST/+orient+3_label *107526 UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 5* AA046634 (1 10)

Each entry (tag annotation) contains five attributes, i.e. the 10-bp tag (bold), the sequence type of the clones from which the tag was extracted (underlined), the UniGene cluster number and cluster name (italic), the accession codes of clones (comma delimited list) and two frequency numbers (between parentheses).

The sequence type provides information about the reliability of the determination of the 3′-end of the GenBank sequence. Since tags are only valid if extracted adjacent to the most 3′ CATG in the sequence, it is very important to establish whether the sequence indeed includes the 3′-end. The following sequence types are defined:

'1/mRNA/+orient'	Well-characterized mRNA or RefSeq sequence (Pruitt and Maglott, 2001).
'2/EST/+orient+3_label'	EST, with polyA signal and/or polyA tail, and labelled as 3′
'3/EST+orient'	EST, with polyA signal and/or polyA tail, but unlabelled
'4/EST+orient+5_label'	EST, with polyA signal and/or polyA tail, and labelled as 5′
'5/EST+3_label'	EST, without polyA signal or polyA tail, but labelled as 3′

The polyA signal and polyA tail both provide information about the 3′-end of the sequence. In the definition of these sequence types only the two most common polyA signals (ATTAAA and AATAAA) were considered. A polyA tail was defined as a stretch of 10 consecutive As at the end of the sequence of 10 consecutive Ts at the beginning of the sequence. Additional information to identify 3′-end sequences is obtained from the depositors of the cDNA sequences, which have assigned a label (3′ or 5′) to the GenBank sequence based on the cloning and sequencing procedures.

The frequency numbers provide information about the reliability and uniqueness of the tag. The first frequency number denotes the number of GenBank clones of this type, with this tag and this UniGene cluster assignment. The second frequency number denotes number of GenBank clones of this type with this tag in any UniGene cluster. In the example above, we see that the tag AAAAATACAA (5/EST+3_label) corresponds to four clones in UniGene cluster 43744. However, from the second frequency number it can be seen that this tag of this type is also extracted from two clones in one or two other UniGene clusters. Therefore, this tag is not unique for a gene or it may be an incorrect tag.

15.4.1 The Construction of Tag-to-Gene Mapping in HTM

To obtain a reliable mapping of gene expression profiles to chromosomes it is important to have a tag-to-gene mapping in which false positive tag identifications (tags that are extracted from the wrong position of the database sequence and therefore do not correspond to the experimentally determined tag of the gene) are removed. False positive tags would strongly compromise the genome-wide expression patterns. The CGAP SAGEmap tag-to-gene mapping contains many false positive tags because this mapping was designed to include all potential tags. To improve the quality of SAGE analysis, the Academic Medical Centre (AMC) tag-to-gene mapping process was constructed to exclude as many false positive tags as possible. The AMC tag-to-gene mapping basically comprises four steps:

1. Identification of the 3′-end of cDNA clones and the electronic extraction of tags.
2. Removal of erroneous tags that result due to EST sequence errors in the 10-bp tag.
3. Removal of erroneous tags that result due to EST sequence errors in the CATG sequence.
4. Identification of anti-sense tags.

Sequencing of cDNA clones occurs, by definition, from the 5′-end to the 3′-end of the sequence. The 5′ → 3′ sequence is called the ‘sense’ sequence, while the 3′ → 5′ sequence is called the ‘complementary-reverse’ sequence. This implies that the most likely orientation of sequences in a database of sequenced cDNA clones is either ‘sense’ or ‘complementary-reverse’. In the case of 3′-end sequences this will, respectively, show the polyA tail as an A-stretch at the end or as a T-stretch at the beginning of the sequence. However, two other possible sequence orientations (reverse or complement) occur in the GenBank database as a result of human errors in submitting or processing the sequence. The frequency of the four possible sequence orientations were analysed by using the 718,271 clones included in the CGAP SAGEmap tag-to-gene mapping of which 12,381 clones contain a stretch of >30 As or Ts at either end of the sequence. Of these clones, 11,476 (93%) end with >30 As (sense) or start with >30 Ts (complementary-reverse). Only 7% of the polyA tails are on the wrong side of the sequence and these clones could result from wrong sequence orientation in the database. Therefore, only the sense and complementary-reverse sequence orientations are considered in the subsequent electronic tag extraction procedures to build the AMC tag-to-gene map. The algorithms that were constructed to build the AMC tag-to-gene map used the cDNA clones (and UniGene cluster assignment) that are included in the CGAP SAGEmap tag-to-gene map. In addition, the sequence type that was assigned to each tag was used.

15.4.2 Identification of 3′-end cDNA Clones and Electronic Tag Extraction

The 3′-end of a processed gene transcript is characterized by a polyA tail and a polyA signal. Besides the two 'classical' polyA signals (AATAAA and ATTAAA), other polyA signals have been reported (Proudfoot, 1991; Sheets *et al*., 1990; see Chapter 12 for more details). The clones included in the CGAP SAGEmap tag-to-gene map were analysed for the occurrence of 'alternative' polyadenylation signals. The clones containing either >30 As at the end or >30 Ts at the beginning of their sequence were selected. Polyadenylation signals are thought to occur within 50 to 100 bp from the polyA addition site (Salamov and Solovyev, 1997). Therefore the 150 nucleotides adjacent to the polyA or polyT stretch were analysed for the presence of the two classical polyadenylation signals, nine possible alternative polyA signals (AATTAA, AATAAC, AATAAT, AATACA, ACTAAA, AGTAAA, CATAAA, GATAAA, TATAAA) and six random hexamer sequences. The two classical polyA signals were found in 55.8 and 17.7% of those clones respectively, and showed a clear preference for occurrence within the first 50 nucleotides from the polyA tail. Four possible alternative polyA signals (AATTAA, AATAAT, CATAAA, AGTAAA) occur in these 50 nucleotides with a frequency ranging from 5.7 to 8.4%. The other five possible polyA signals and the six random hexamers showed no appreciable preference for occurring in the 3′-end of transcripts. Therefore, the sequence orientation algorithms that were developed were configured to search for the six most abundant polyA signals within 50 bp from the polyA site. The same frequency and position patterns for the six polyA signals were found in cDNA clones ending with at least 10 As or starting with at least 10 Ts. This indicates that the occurrence of stretches of 10 or more As or Ts at the end and the beginning of a cDNA sequence, respectively, is likely to represent a polyA tail.

The sequence types that are included in the CGAP SAGEmap tag-to-gene map provide additional information to identify the 3′-end clones. This sequence type was combined with the presence of one of the six polyA signals at either end of the clone sequence (within 50 bp) and/or a polyA tail (>10 As at the end or >10 Ts at the beginning) to select for reliable 3′-end clones. To minimize the risk of extracting erroneous tags (false positives) from GenBank sequences, only 'reliable 3′-end' clones were used for electronic tag extractions. When both strands of a cDNA encoded conflicting polyadenylation signals and/or polyA/polyT stretches, clones were not used for tag extraction.

15.4.3 Identification of 10-base Pair Tag Sequencing Errors

Single pass high throughput sequencing of EST libraries is one of the more error prone sequencing methods; therefore the chance of a sequence error is about 1% per base. Consequently, tags that are electronically extracted from database sequences may include sequencing errors. Therefore, the tags were checked for errors in the 10-bp sequence resulting from sequencing errors in ESTs. If it is assumed that sequencing errors are independent for each base and the error rate is 1%, then the probability of one error being present is only $10 \times 0.01 \times 0.99^9 = 0.091$. We designed algorithms that detected any combination of matching tags with maximal two-base substitutions, insertions or deletions because the chance that a tag will contain three errors is negligible (0.01%). To check for sequencing errors all EST clones in a UniGene cluster were compared pair-wise and checked for substitutions, insertions or deletions. If two tags were identical, except for one or two mismatches, a potential sequencing error in the tag might be involved. The tag corresponding to the largest number of clones was considered to be a correct

tag. The tag with the potential sequencing error was removed when it was found in less than five ESTs/cDNAs and was five times less frequent then the correct tag. This ensured that variant tags resulting from frequent single nucleotide polymorphisms (SNPs) were not discarded in the AMC tag-to-gene mapping.

15.4.4 Identification of CATG Sequencing Errors

Sequence errors (Figure 15.2) in the most 3′ CATG sequence of an EST will result in skipping of the corresponding tag by the extraction algorithm and erroneous use of the next CATG for tag extraction. Also, an EST sequence error may create a new CATG distal

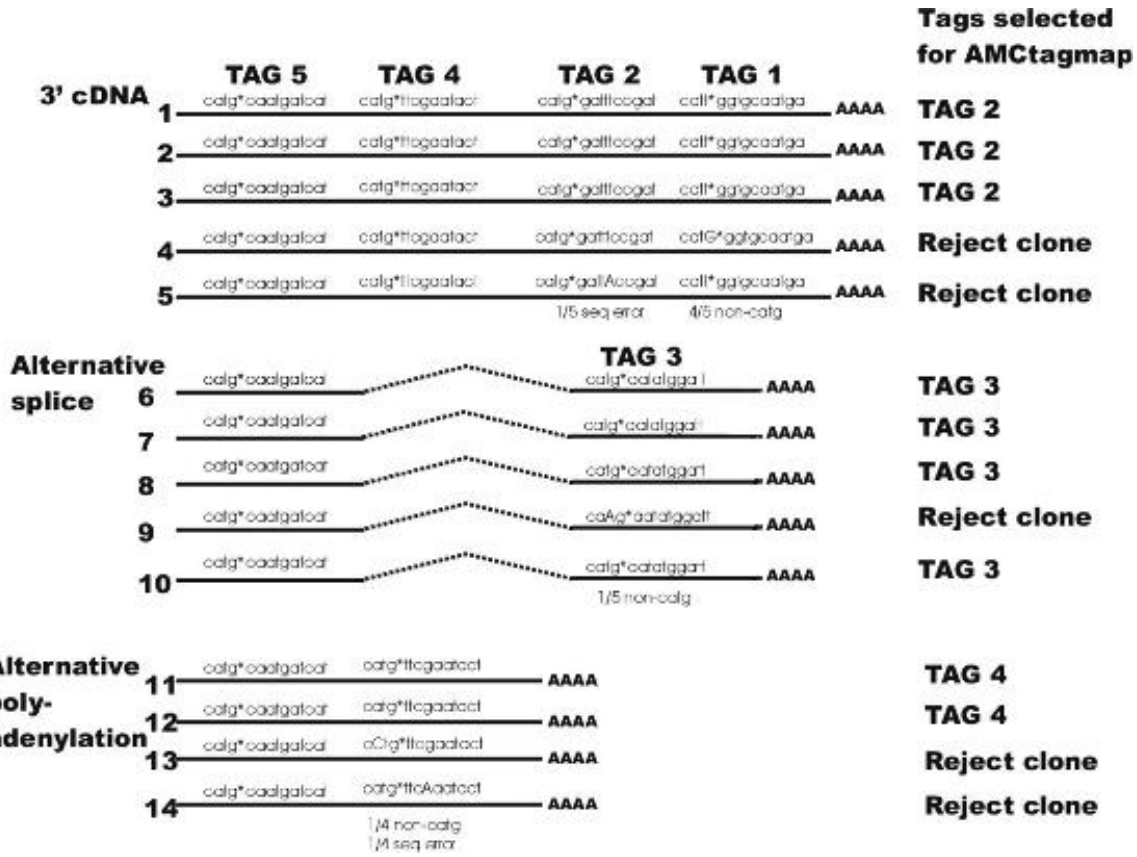


Figure 15.2 Identification of (CATG) sequencing errors. This example shows 15 EST clones (five 3′ cDNA clones, five 3′ cDNA clones of the alternatively spliced gene and five 3′ cDNA clones of the alternatively polyadenylated gene). TAG2 (GATTTCCGAT) is the correct tag for the first five clones. However, clone 4 is rejected because a CATG is created due to a sequencing error (T → G). If this clone was not rejected then TAG1 (GGTGCAATGA) would mistakenly be associated to this transcript. Clone 5 is rejected because TAG2 contains a sequencing error. Both sequencing errors are not considered to be SNPs because they only occur once in these five clones. TAG5 (AATATGGATT) is the correct tag for the alternatively spliced gene. Clone 9 is rejected because the CATG is destroyed due to a sequencing error (T → A). If this clone was not rejected then TAG5 would be mistakenly associated to this clone. In the case of the alternatively polyadenylated genes no clones are rejected because too few clones are available to make a decision. Consequently, TAG4 (TTCGAATACT) is extracted from clones 11 and 12, TAG4 (CAATGATCAT) from clone 13 (CATG was destroyed) and TAG6 (TTCAAATACT) from clones 14 and 15.

to the true most 3′ CATG. This also results in extraction of a false tag for an EST. An algorithm to remove these tags should preserve tags from alternatively spliced transcripts of the same gene. Each gene can have a series of tags belonging to alternatively spliced or alternatively polyadenylated transcripts. Furthermore, SNPs in the CATG sequence can cause extraction of alternative tags that are correct and should be preserved. Our algorithms were directed to the identification and removal of all tags that are caused by CATG sequence errors. The remaining tags were accepted as reliable tags.

15.4.5 Identification of Sense and Antisense Tags

One of the major problems with the UniGene clustering algorithm is that it can place overlapping genes encoded on opposite DNA strands in one UniGene cluster. In such cases, tag extraction routines may extract the tags from both genes. Therefore, algorithms to recognize oppositely oriented tags were designed. In such clusters, the orientation of the most frequent tag was considered as 'sense'. The antisense tags were marked and preserved in the AMC tag-to-gene mapping.

15.4.6 Comparison of SAGE Libraries

The HTM does not include statistical routines to establish whether two expression levels are significantly different. Therefore, once a candidate gene has been identified (based

TABLE 15.1 Public Resources (Software and Databases) Available for the (Statistical) Analysis of SAGE Data

Resource	Main Functionalities	Website
SAGE300 (Zhang *et al.*, 1997)	Tag extraction, tag identification, statistical comparison	www.sagenet.org
CGAP SAGEmap (Lal *et al.*, 1999)	Tag identification, statistical, xProfiler, Virtual Northern	www.ncbi.nlm.nih.gov/SAGE/
USAGE (van Kampen *et al.*, 2000)	Tag extraction, tag identification, statistical comparison, management of SAGE libraries (pool, merge, etc.)	www.cmbi.kun.nl/usage/
eSAGE (Margulies and Innis, 2000)	Tag extraction, statistical comparison, data management	ehm@umich.edu
Detecting sequencing errors (Colinge and Feger, 2001)	Detection of sequencing errors in SAGE libraries	georg.feger@serono.com
Audic and Claverie (1997)	Statistical comparison	igs-server.cnrs-mrs.fr/□audic/significance.html
SAGEstat (Kal *et al.*, 1999)	Statistical comparison	j.m.ruijter@amc.uva.nl or www.cmbi.kun.nl/usage/
POWER_SAGE (Man *et al.*, 2000)	Statistical comparison	michael.man@pfizer.com

on visual inspections of tag counts) one may calculate the statistical difference between the tag counts. Several statistical methods are available (see also Table 15.1) and are discussed in this section.

The aim of statistical comparison of two SAGE libraries is to reject the null hypothesis that the observed tag counts in both libraries are equal. Testing of this hypothesis is hampered by the fact that SAGE experiments are generally not repeated, and therefore, each SAGE library is only one measurement: the necessary information on biological variation and experimental precision is not available in the data. It is possible that all differences between two libraries are just the result of random sampling from the same population. Therefore, before starting a pair-wise comparison of specific tags in two libraries, the null hypothesis that the differences between libraries result from random sampling has to be rejected. In the context of SAGE research, only one reference to a test for this purpose has been published (Michiels *et al.*, 1999). This overall test is based on a simulation of a large number of possible distributions of two libraries within the pooled marginal totals of the observed SAGE libraries. By calculating the chi-squared statistic for each simulated pair of libraries, a distribution of this statistic under the null hypothesis can be constructed. From this simulated distribution and the chi-squared value of the observed libraries, one can then determine the probability of obtaining the observed tag distributions at random. Rejection of the null hypothesis that all differences between SAGE libraries are just the result of random sampling then opens the way for pair-wise comparisons.

15.4.7 Statistical Tests for Differences Between SAGE Libraries

Several statistical tests have been published for the pair-wise comparison of SAGE libraries. For all tests the null hypothesis states that there is no difference in tag numbers between the two libraries that are compared. It should be kept in mind that in most comparisons between specific tags in SAGE libraries, there is no *a-priori* knowledge about the direction of the effect. Therefore, all decision rules have to be formulated to result in a two-sided test. The significance level (α) can be set to 0.001 to safeguard against the rate of accumulation of false positives that may result from multiple testing (Bonferroni correction; Altman, 1991).

The different methods that can be used to test the difference between two SAGE libraries can be compared by considering the critical values. Critical values are defined as the highest or lowest number of tags that, given an observed number of tags in one library, needs to be found in the other library to result in a p-value below the significance level when the pair-wise test is carried out. They can be determined by repeatedly testing simulated tag numbers until the resulting p-value leads to rejection of the null hypothesis at the required level of significance.

In the original SAGE paper (Velculescu *et al.*, 1995), tag numbers in different libraries are compared pair-wise with a test based on a Monte Carlo simulation of tag counts. This approach is included in the SAGE software package SAGE300 (Zhang *et al.*, 1997). SAGE300 performs, in each pair-wise comparison, at least 100 with a maximum of 100,000 simulations to determine the chance of obtaining a difference in tag counts equal to or greater than the observed difference. This results in a one-sided p-value that has to be compared to $\alpha/2$. Since the Monte Carlo-based test of SAGE300 does not give the same p-value every time the same input is tested, each input is run six times and the mean p-value is used for the determination of the upper critical values that are given in Figure 15.3A. In this figure the critical values are given for two SAGE libraries of equal size (diamonds) and for two SAGE libraries of different size (squares). The critical

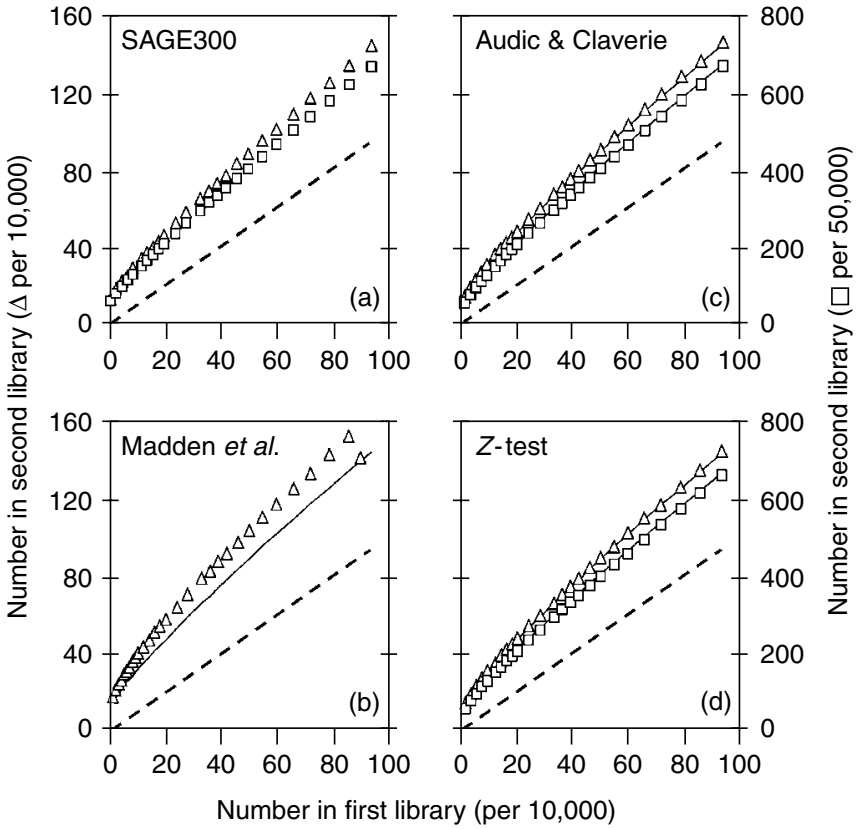


Figure 15.3 Comparison of the critical values of different tests for SAGE data. Critical values are defined as the numbers of tags that need to be found in the second SAGE library to be significantly different from the number of tags already found in the first SAGE library. Upper critical values for a 0.001 level of significance are given for (A) SAGE300 (Zhang *et al*., 1997), and the tests of (B) Madden *et al*. (1997), (C) Audic and Claverie (1997) and (D) the Z-test of Kal *et al*. (1999). The critical values plotted in each graph are based on a first SAGE library with a total of 10,000 tags (reference values, plotted as a dotted continuous line on the x-axis) and a second library with a total of 10,000 tags (critical values plotted as triangles on the left y-axis) or a second library of 50,000 tags (critical values plotted as squares on the right y-axis). In B, C and D a plot of the critical values of SAGE300 (A) are added (thin lines) to facilitate comparison between tests. In B only critical values for a second library of 10,000 tags are given because Madden's test can only be used for libraries of similar size.

values of SAGE300 are copied as continuous lines into Figure 15.3B, C and D to facilitate comparison with other tests.

The test suggested by Madden *et al*. (1997) is based on only the number of observed specific tags in each SAGE library and the test statistic is calculated as:

$$Z = \frac{n_1 - n_2}{\sqrt{n_1} + \sqrt{n_2}} \tag{1}$$

with n_1 and n_2 as the number of specific tags in the first and second library, respectively. This test statistic is estimated to be normally distributed and can be compared to $Z_{\alpha/2}$. The test of Madden requires about 25% larger differences than SAGE300 to reach statistical significance and is, therefore, more conservative (Figure 15.3B). Only one set of critical

values is given because this test can only be used for two libraries of similar size. However, the simple mathematics of this test (Eq. 1) are a point in its favour.

Audic and Claverie (1997) derived a new equation for the probability of finding n_2 or more tags in one library given the fact that n_1 tags have already been observed in the other library:

$$P(n_2|n_1) = \left(\frac{N_2}{N_1}\right)^{n_2} \frac{(n_1 + n_2)!}{n_1!n_2!(1 + N_2/N_1)^{(n_1+n_2+1)}} \tag{2}$$

with N_1 and N_2 as the total number of tags in the first and second library, respectively. A summation of this probability over all n from n_2 to infinity gives a one-sided p-value that can be compared to $\alpha/2$. The upper critical values for a significance level of 0.001 for Audic and Claverie's test are given in Figure 15.3C. For both the libraries of equal and different size these critical values are all within 1.5% of those of SAGE300.

The Z-test focuses on the proportions of specific tags in each library and is based on the normal approximation of the binomial distribution (Altman, 1991; Kal *et al*., 1999). The test statistic Z is calculated as the difference in proportions divided by the standard error of this difference:

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1 - p_0)(1/N_1 + 1/N_2)}} \tag{3}$$

with $p_1 = n_1/N_1$ and $p_2 = n_2/N_2$. The proportion p_0, the expected proportion when the null hypothesis is true, is calculated as $p_0 = (n_1 + n_2)/(N_1 + N_2)$. Z is approximately normally distributed and can be compared to $Z_{\alpha/2}$. The critical values of the Z-test are given in Figure 15.3D and are also all within 1.5% of those of SAGE300.

The chi-squared test can be used for comparing SAGE libraries (Michiels *et al*., 1999) after reorganizing the data in a 2 × 2 contingency table. However, this test is statistically equivalent to the Z-test on two proportions (Altman, 1991) and will give the same p-values and have the same critical values. Another test using 2 × 2 contingency tables is the Fischer exact test (Altman, 1991), which has also been applied to SAGE data (Man *et al*., 2000). However, the sampling design required by this test does not apply to SAGE (Claverie, 1999; Conover, 1980) and moreover, for the large number of tags involved in SAGE, the chi-squared test is to be preferred. In the paper by Chen *et al*. (1998), a procedure based on Bayesian statistics is described to calculate the probability that the level of expression of a given mRNA is increased by at least x-fold between libraries. Although this procedure can be used to statistically judge differences in tag numbers, its approach is clearly different from the classical approach of hypothesis testing and results of these test procedures cannot be directly compared.

In conclusion, this comparison shows that SAGE300, Audic and Claverie's test (1997) and the Z-test, will all give the same test results when applied for pair-wise comparison of SAGE libraries whereas Madden's test will behave considerably more conservatively. In a Monte Carlo comparison of the chi-squared test, Fischer exact test and Audic and Claverie's test it was shown that the chi-squared test, which is equivalent to the Z-test, had the best power and robustness (Man *et al*., 2000), especially at low expression levels.

15.4.8 Computational Resources for SAGE Analysis

Table 15.1 summarizes the public resources that are available for the analysis of SAGE data. The SAGE300 program (Zhang *et al*., 1997) is probably the most commonly used application for SAGE analysis. To identify SAGE tags the SAGE300 program compiles a

tag-to-gene map from human (EST) sequences in GenBank. A drawback of this method is that the orientation of the sequence is not checked before tag extraction and consequently, incorrect tags can result. SAGE300 also includes a Monte Carlo-based method for statistical comparison of SAGE libraries.

As part of CGAP the NCBI established the SAGEmap public database (Lal *et al*., 1999), which includes SAGE libraries and a tag-to-gene mapping. SAGEmap also includes a 'reliable tag-to-gene map', which accounts for sequencing errors in GenBank sequences. These tag-to-gene maps can be downloaded and used in combination with applications such as Microsoft Access. Alternatively, the tag-to-gene maps are accessible online from the SAGEmap site but this only allows the analysis of one tag at a time. No full identification reports, i.e. for all tags in a SAGE tag list, can be generated as is possible with SAGE300, which unfortunately does not support the use of these tag-to-gene maps.

The USAGE application (van Kampen *et al*., 2000) allows construction of tag-to-gene maps from the EMBL database for any organism. The program allows the extraction of tags from the sense and complement-reverse orientation of the sequence because the 3′-end of the clone is not determined prior to tag extraction. However, USAGE also includes both SAGEmap tag-to-gene maps and the AMC tag-to-gene map and allows the user to produce full tag identification reports. USAGE includes the *Z*-test for the statistical comparison of SAGE libraries (Kal *et al*., 1999).

The eSAGE software (Margulies and Innis, 2000) is similar to USAGE. It includes the SAGEmap tag-to-gene mapping and performs statistical comparisons according to the test proposed by Claverie (1999). The input concatemers can contain any characters from the standard IUPAC code. In addition, eSAGE reads PHD files generated from phred-analysed sequence trace files (Ewing and Green, 1998; Ewing *et al*., 1998) and uses the phred quality values for each base as a more accurate method of excluding low quality sequence data.

Colinge and Feger (2001) introduced a method to identify possible sequence errors in tags in SAGE libraries. This method in combination with an accurate tag-to-gene map can greatly enhance SAGE tag identification.

15.5 INTEGRATION OF BIOLOGICAL DATABASES FOR THE CONSTRUCTION OF THE HTM

To enable the mapping of gene expression profiles to chromosomes in the HTM, several public databases were integrated in a relational database. The HTM was constructed by mapping gene expression levels (SAGE tag counts) to gene positions as defined by the GeneMap99 database (Deloukas *et al*., 1998). GeneMap99 gives the chromosomal position of 45,049 human expressed sequence tags (ESTs) and genes belonging to 24,106 UniGene clusters. The STS markers in GeneMap99 are assigned to a unique radiation hybrid code (RH-code), which is linked to the accession code of the corresponding clone in the rhdb_xrefs_human cross-reference file, which is part of the radiation hybrid database (RHdb; Rodriguez-Tome and Lijnzaad, 1997). This accession code is linked to the AMC tag-to-gene mapping to obtain the corresponding UniGene cluster and thereby the corresponding SAGE tags. The tags from the tag-to-gene mapping are linked to the expression levels in the selected SAGE libraries. If an accession code of an STS marker was not present in the cross-reference file then the UniGene cluster was retrieved instead of the accession code. The UniGene cluster was then used to retrieve the corresponding SAGE tags in the tag-to-gene map and the expression levels in the SAGE libraries.

15.5.1 The HTM Relational Database

E. F. Codd at IBM introduced the relational database in 1970 (Codd, 1970), since then this form of database has developed to fundamentally underpin most modern bioinformatics databases. A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to re-organize the database tables (Ullman, 1988). Each table contains one or more data categories in columns. Each row contains a unique instance of data for the categories defined by the columns. It is important to carefully design the database model because a poorly designed database may be slow to query, hard to maintain and extend, and may contain inconsistent and redundant information.

15.5.2 Relational Database Design

Relational databases are a key concept in bioinformatics and so it is useful to take the HTM as an example of database design and construction. The integration of the aforementioned public databases and SAGE libraries into the HTM relational database is shown in an entity–relationship (ER) diagram (Figure 15.4). The ER diagram describes the HTM

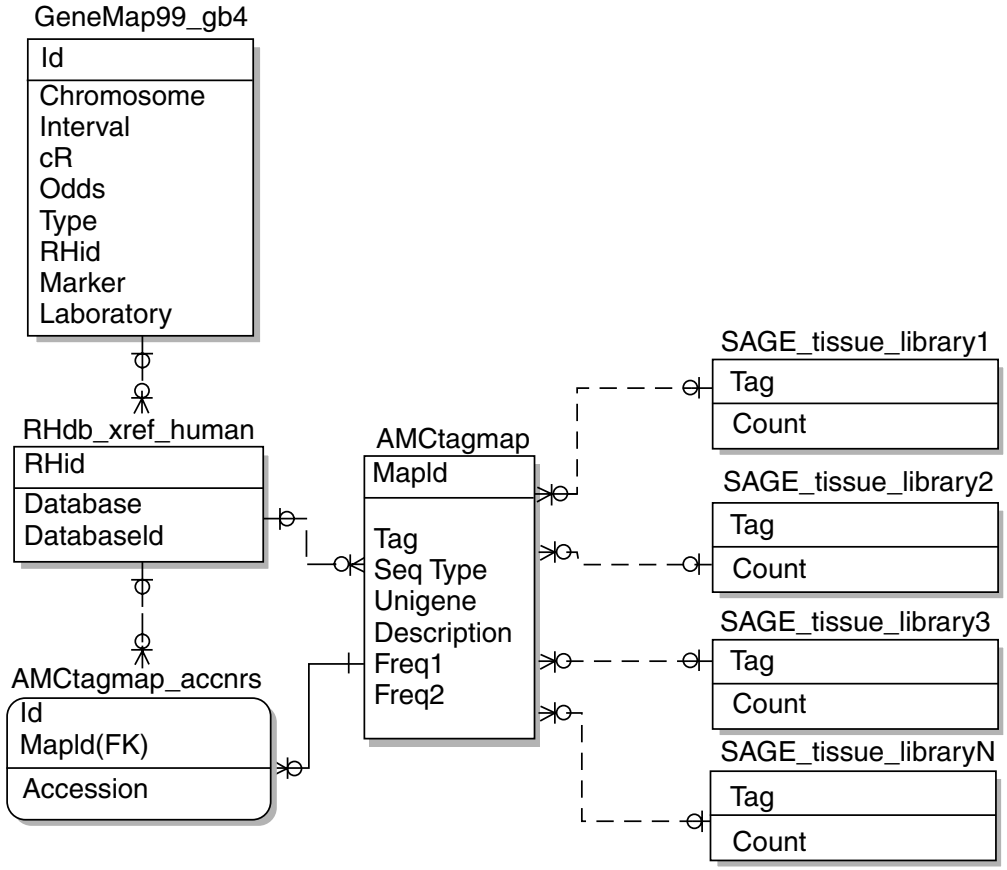


Figure 15.4 Relational model of database used in the HTM. Each table in the database (e.g. AMCtagmap) contains a number of attributes (e.g. Tag). The relationship between the tables are specified as 'zero-or-one to many' or as 'one to many'. For example, each tag in a SAGE library is linked to zero or more electronic tags in the 'AMCtagmap' table. Subsequently, each of these tags is linked via the 'RHdb_xref_human' table to the 'GeneMap99_gb 4' table to establish the mapping.

database tables and relationship between these tables. The relational database model was implemented by using the Postgresql relational database management system (RDBMS) (http://www.postgresql.org). A RDBMS tool allows the developer to:

1. Implement a database with tables, columns and indexes.
2. Define the so-called foreign keys, which specify relationships between rows of various tables.
3. Update the indexes automatically.
4. Interpret an SQL query and combine information from various tables.

Once the tables are implemented it is possible to upload data to the database or extract data from the database by using SQL (Structured Query Language). SQL is the standard user and application program interface to a relational database and is used both for interactive queries for information from a relational database and for gathering data for reports. The reader should be aware that despite first impressions, SQL is a very easy language to learn; 2 days' training can quickly enable a new user to perform complex database queries to integrate diverse forms of data. For example, the next SQL query returns all expression levels of genes mapped on chromosome 1 (compare the statements in this query with the ER diagram in Figure 15.4 to get an idea of what this query is doing):

```
SELECT gm.Chromosome, gm.cR, amc.Unigene, SUM(sage.Count)
FROM GeneMap99_gb4 AS gm JOIN RHdb_xrefs_human  AS rh
ON (gm.RHid = rh.RHid) JOIN AMCtagmap  AS amc
ON (rh.Databaseid = amc.Unigene) JOIN SAGE_tissue_library1
    AS sage
ON (amc.Tag = sage.Tag)
WHERE gm.chromosome = 'chr1'
AND rh.DatabaseName = 'UniGene'
GROUP BY gm.Chromosome, gm.cR, amc.Unigene
ORDER BY gm.cR
```

The relational database forms the core of HTM in which all required data to map expression profiles to chromosomal positions are stored. The SQL queries are part of the user-interface that is built on top of the relational database and which is introduced in the next section.

15.6 THE HUMAN TRANSCRIPTOME MAP

The Human Transcriptome Map (HTM; bioinfo.amc.uva.nl) is a database application that presents gene expression profiles for any chromosomal region in normal and pathological tissues (Caron *et al.*, 2001). The application can be used to search for genes that are over-expressed or silenced in cancer. The HTM provides three different ways to present gene expression profiles obtained with SAGE. The 'extended view' provides the most detailed level of information (Figure 15.5). In this view the expression profiles given for all SAGE tags that could be linked to the radiation hybrid map (RH-map) are shown. Different tags may correspond to a single gene as they may occur as a result of differential splicing or polyadenylation of the gene. In the 'concise view', no individual tags are included but information is presented at the gene (UniGene) level and consequently the tag counts for all tags belonging to the same genes are pooled, i.e. no distinction is

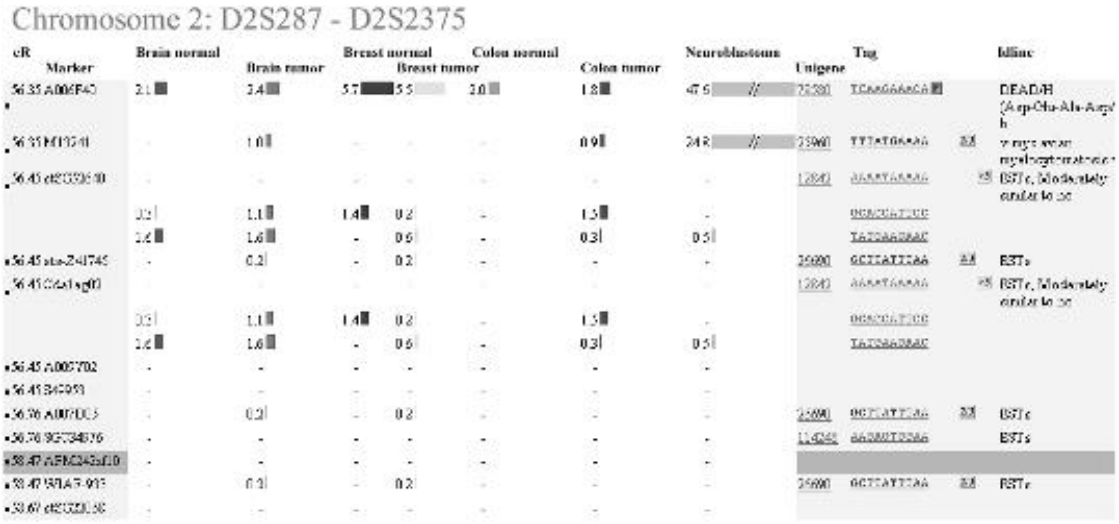


Figure 15.5 Extended view of a chromosome 2p region showing neuroblastoma-specific over-expression of the neighbouring genes N-myc (UniGene Hs.25960) and DDX-1 (UniGene Hs.78580). A small part of the interval D2S287 to D2S2375 is shown. The left-hand columns show the marker and centiRay position as defined on GeneMap99. The right-hand side shows the UniGene number, tag sequence and the description of the UniGene cluster. Expression levels in the libraries are normalized per 100,000 tags and shown by grey bars with a range from 0 to 15. Numbers give the counts per 100,000 tags. The tags are annotated by symbols (explained in the text). (*Reprinted with permission from Caron et al. (2001). Copyright 2001 American Association for the Advancement of Science*).

made between different gene variants. In both the concise and extended view, only a selected region between two framework markers of a chromosome is shown. In the 'whole chromosome view' the expression levels of all genes on a particular chromosome are displayed (Figure 15.6). Also in the whole chromosome view the tag counts for all tags belonging to the same gene are pooled to obtain an overall expression level. In this presentation each unit on the vertical axis represents one gene, i.e. the scale does not denote a genetic or physical distance. The RH-map contains errors (see Chapter 7) and, therefore, some genes map two or more times at slightly different positions. Genes that correspond to multiple markers on the RH-map are shown only on the HTM at the position of the highest LOD score. Only genes for which a tag was included in the AMC tag-to-gene map are displayed.

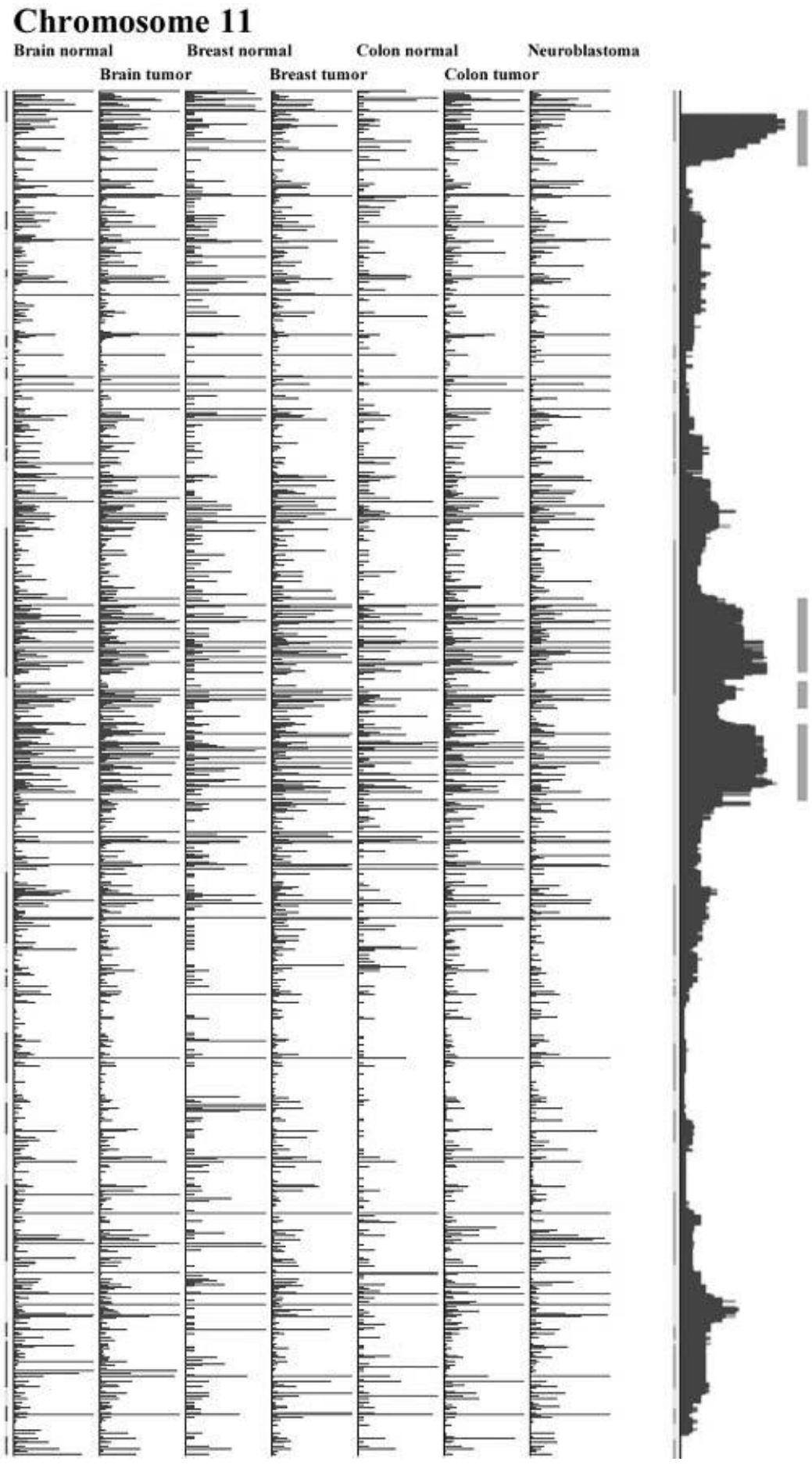
15.6.1 Annotation of the HTM

In the extended and concise view of the HTM, several annotation symbols are used.

15.6.1.1 Unreliable Tags

Two types of tags were considered unreliable for use in the HTM. They are marked as 'L', '2/3' or '>3' in a yellow box:

1. **Linker tags**. The SAGE technique may produce tags derived from linker oligo's used in library construction (V. E. Velculescu *et al*., personal communication). These 73 linker tags are marked 'L' in a yellow box on the extended interval view, but their expression levels in the SAGE libraries are not shown.


Chromosome 11
Brain normal
Brain tumor
Breast normal
Breast tumor
Colon normal
Colon tumor
Neuroblastoma

2. **Redundant tags**. Some tags are found for more than three UniGene clusters. This may be explained by coincidental limited sequence homologies between genes. Other redundant tags are derived from genes with a CATG close to the polyA tail. This generates tags with a strongly reduced sequence variability, as most of the tag consists of an A stretch. They are marked '>3' in a yellow box in the extended interval view and their expression levels in the SAGE libraries are *not* shown. Tags belonging to two or three UniGene clusters are marked in a yellow box with '2/3' respectively, and their expression levels in the libraries are shown.

15.6.1.2 Antisense Tags

In the extended interval view, tags with an antisense orientation are marked as 'AS' in a purple box. In the concise interval view, the cumulative expression levels for 'sense' and 'antisense' tags are shown as separate bars for each UniGene cluster. Antisense expression levels are not included in the whole chromosome views.

15.6.2 UniGene Clustering Errors

Hybrid UniGene clusters cause many problems, as they include ESTs from different genes. These genes, which usually have different map positions, each yield their own correct reliable tags. To identify the hybrid clusters the GenBank database (Genomes *Homo sapiens* section) was searched for the corresponding PAC sequenced in the Human Genome Project, as well as two adjacent PACs, for the markers mapped on GeneMap99. Tags from the gene corresponding to the marker are expected to be present on these PACs, whereas tags from a 'contaminating' gene in a hybrid cluster are not. The PACs were analysed for the presence of the 10-bp tag sequence plus adjacent CATG. When positive, the tag was marked on the extended interval view with a 'P' in a light green box. A one-nucleotide mismatch between tag and PAC sequence was accepted to cover SNPs or PAC sequencing errors (marked 'P' in a dark green box). When a PAC for a marker was known, but when the tag was not found in the sequence, the tag was marked 'P' in a red box. For all situations the expression level of the tag is shown in all views. This check is not yet available for all markers, but the progress in sequencing and annotation will provide this function for all UniGene clusters.

15.7 REGIONS OF INCREASED GENE EXPRESSION (RIDGES)

The Human Transcriptome Map provides an intriguing insight into the higher-order organization and regulation of expression in the human genome. From the whole chromosome views it is clear that there is a strong clustering of highly expressed genes in specific

---

Figure 15.6 Whole chromosome view of expression levels of the 1208 UniGene clusters mapped to chromosome 11 on the GB4 radiation hybrid map of GeneMap99. Each unit on the vertical axis represents one UniGene cluster. Expression is shown for SAGE libraries of 7 out of the 12 available tissue types. Expression levels in the libraries are normalized per 100.000 tags and tag counts from 0 to 15 are shown by horizontal blue bars while tag frequencies over 15 are shown as red bars (colors not shown in this figure). The section to the right represents a moving median with a window size of 39 UniGene clusters generated from the expression levels in 'all tissues'. The bars above the moving median indicate RIDGEs. (*Reprinted with permission from Caron et al. (2001). Copyright 2001 American Association for the Advancement of Science*).

Figure 15.7 Comparison of median gene expression levels and gene density for chromosome 3. The lower diagram shows the expression levels as a moving median with a window size of 39 UniGene clusters. The upper diagram shows gene density. For each UniGene cluster, the average distance between adjacent clusters in a window of 39 adjacent UniGene clusters was calculated. The inverse of this value is shown (inverse centiRays per gene). (*Reprinted with permission from Caron et al. (2001). Copyright 2001 American Association for the Advancement of Science*).

domains, which were named Regions of Increased Gene Expression (RIDGEs) (Caron *et al*., 2001). This is clearly demonstrated in Figure 15.6, which shows the whole chromosome view of expression levels of 1208 genes mapped to the RH-map of chromosome 11. Expression is shown for SAGE libraries of seven tissue types. To emphasize the RIDGEs more clearly, a moving median with a window size of 39 genes was calculated for 'all tissues', which pools all available SAGE libraries. From the resulting median values, RIDGEs were defined as regions in which at least 10 consecutive genes have a median expression level of at least four times the genomic median. Green bars in the resulting graph indicate the resulting RIDGEs. These RIDGEs were observed on most chromosomes. With the current definition, 27 RIDGEs could be identified (Caron *et al*., 2001).

Analysis of RIDGEs for physical characteristics suggests that many of them have a high gene density. Figure 15.7 shows the correlation between RIDGEs and gene density (expressed as cR^{-1}/gene) for chromosome 3. This correlation between gene expression and density of mapped genes is found for most RIDGEs. Typical RIDGEs contain six to 30 mapped genes per centiRay, compared to one to two mapped genes per centiRay for weakly transcribed regions.

15.7.1 Statistical Evaluation of RIDGEs

To analyse whether the observed RIDGEs could be explained by the random variation in the distribution of expression levels of the 18,422 UniGene clusters in the HTM, a Monte Carlo simulation was performed. We permutated the genomic order of all 18,422 UniGene clusters in the Human Transcriptome Map and analysed 10,000 permutated datasets for the incidence of RIDGEs. The number of RIDGEs according to our definition was determined for each of the permutations. The observed number of RIDGEs in the Human Transcriptome Map (27) was about 38 standard deviations (0.7) higher than the average number of RIDGEs (0.4) observed in the permutations. The observed number of RIDGEs is therefore unlikely to result from random variation in the distribution of highly expressed genes over the genome.

15.8 DISCUSSION

This chapter has reviewed one possible approach to the analysis of gene expression data in which (statistical) data analysis, database technology, informatics and molecular biology

play an important role. The HTM was designed to assist in the identification of genes that are involved in cancer; however it has a wider applicability to the study of any disease. In Chapter 9, approaches for the expression-based prioritization of positional gene candidates in disease loci were reviewed. The HTM could be a valuable tool for prioritizing such candidates. As the number of publicly available SAGE libraries increases their value will also increase (as every SAGE experiment can be directly compared).

Perhaps the most interesting aspect of the SAGE-based, Human Transcriptome Map is that it is somewhat different from other approaches and therefore it is complementary to microarray data. Integration of (public) databases using HTM, uncovered a previously unknown genomic phenomenon — regions of increased gene expression (RIDGEs). RIDGEs may provide more fundamental insight into the higher-order organization of the human genome. The biology of RIDGEs is not yet understood but they may play an important role in gene transcription and therefore, may be relevant to the study of carcinogenesis or any other disease which involves disregulation of gene expression.

RIDGEs would not have been revealed if DNA microarray data had been used. Since the overall expression profile for all chromosomes is similar in all tissues, the measurement of the expression of one tissue relative to a control tissue would reveal only genes that are differentially expressed between these tissues. Furthermore, as explained in Section 15.2, the expression levels of genes on one DNA microarray cannot be compared and therefore, these domains would not have this clear structure. However, DNA microarray data can be used to further understand the nature of RIDGEs. It can be envisioned that specific tumour samples have disturbed expression of entire transcriptional domains due to translocations. DNA microarrays are very suitable for measuring gene expression profiles for large numbers of (tumour) samples; integration of this data with the HTM would directly reveal whether gene expression in specific domains is turned on or off. Such experiments may further increase our knowledge about the organization of the genome with respect to gene expression.

The current HTM is not the end of gene expression analysis but can be regarded as the starting point of much more research that aims at understanding the biology of RIDGEs. This research includes the construction of a sequence-based HTM that is much more precise than the current map that is based on radiation hybrid data. Such a sequence-based map would allow a more precise definition of RIDGEs. Furthermore, this will allow the investigation of the correlation between RIDGEs and other domains such as gene density. To understand why many genes in RIDGEs are highly expressed in comparison to other regions one could search for regulatory sequences that are common for genes in such domains. Moreover, and maybe more interesting, is the hunt for regulatory sequences that turn complete domains of genes on and off. To enhance the search for regulatory sequences a comparison between the Human Transcriptome Map and a Mouse Transcriptome Map would be very valuable since conserved sequences can be identified (see Chapter 12 for an overview of some of the tools which may be suitable for such an analysis). For all this research much more bioinformatics and laboratory work is required. However, this will ultimately lead to a further understanding of the molecular biology of cancer and human disease.

REFERENCES

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al*. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**: 503–511.

Altman DG. (1991). *Practical Statistics for Medical Research*. Chapman-Hall: London.

Audic S, Claverie JM. (1997). The significance of digital gene expression profiles. *Genome Res* **7**: 986–995.

Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z. (2000). Tissue classification with gene expression profiles. *J Comput Biol* **7**: 559–583.

Boguski MS, Lowe TM, Tolstoshev CM. (1993). dbEST — database for 'expressed sequence tags'. *Nature Genet* **4**: 332–333.

Bowtell DD. (1999). Options available — from start to finish — for obtaining expression data by microarray. *Nature Genet* **21**: 25–32.

Brodeur GM, Sekhon G, Goldstein MN. (1977). Chromosomal aberrations in human neuroblastoma. *Cancer* **40**: 2256–2263.

Brown PO, Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genet* **21**: 33–37.

Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, *et al*. (2001). The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**: 1289–1292.

Chen H, Centola M, Altschul SF, Metzger H. (1998). *J Exp Med* **188**: 1657–1668.

Claverie JM. (1999). Characterization of gene expression in resting and activated mast cells. *Hum Mol Genet* **8**: 1821–1832.

Codd EF. (1970). *Commun ACM* **13**: 377–387.

Cole KA, Krizman DB, Emmert-Buck MR. (1999). The genetics of cancer — a 3D model. *Nature Genet* **21**: 38–41.

Colinge J, Feger G. (2001). Detecting the impact of sequencing errors on SAGE data. *Bioinformatics* **17**: 840–842.

Conover WJ. (1980). *Practical Nonparametric Statistics*, John Wiley: New York.

Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, *et al*. (1998). A physical map of 30,000 human genes. *Science* **282**: 744–746.

Ewing B, Green P. (1998). Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* **8**: 186–194.

Ewing B, Hillier L, Wendl MC, Green P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* **8**: 175–185.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, *et al*. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**: 531–537.

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, *et al*. (2000). 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* **1**: research 0003.1–0003.21.

Heid CA, Stevens J, Livak KJ, Williams PM. (1996). Real time quantitative PCR. *Genome Methods* **6**: 986–994.

Kal AJ, van Zonneveld AJ, Benes V, van den Berg M, Koerkamp MG, Albermann K, *et al*. (1999). Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell* **10**: 1859–1872.

van Kampen AH, van Schaik BD, Pauws E, Michiels EM, Ruijter JM, Caron HN, *et al*. (2000). USAGE: a web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics* **16**: 899–905.

Kanehisa M, Goto S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**: 27–30.

Karp PD. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics* **14**: 753–754.

Karp PD, Paley S, Zhu J. (2001). Database verification studies of SWISS-PROT and GenBank. *Bioinformatics* **17**: 526–532; discussion 533–534.

King RJB. (2000). *Cancer Biology*. Pearson Education: Harlow, UK.

Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, *et al*. (1999). A public database for gene expression in human cancers. *Cancer Res* **59**: 5403–5407.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, *et al*. (2000). SAGEmap: a public gene expression resource. *Genome Res* **10**: 1051–1060.

van Limpt V, Chan A, Caron H, Sluis PV, Boon K, Hermus MC, *et al*. (2000). SAGE analysis of neuroblastoma reveals a high expression of the human homologue of the Drosophila Delta gene. *Med Pediatr Oncol* **35**: 554–558.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, *et al*. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol* **14**: 1675–1680.

Madden SL, Galella EA, Zhu J, Bertelsen AH, Beaudry GA. (1997). SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* **15**: 1079–1085.

Man MZ, Wang X, Wang Y. (2000). POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* **16**: 953–959.

Margulies EH, Innis JW. (2000). eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics* **16**: 650–651.

Michiels EM, Oussoren E, Van Groenigen M, Pauws E, Bossuyt PM, Voute PA, *et al*. (1999). Genes differentially expressed in medulloblastoma and fetal brain. *Physiol Genomics* **1**: 83–91.

Mitelman F, Johansson B, Mertens F. (2001). http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Okamoto H, Yonemori F, Wakitani K, Minowa T, Maeda K, Shinkai H. (2000). A cholesteryl ester transfer protein inhibitor attenuates atherosclerosis in rabbits. *Nature* **406**: 203–207.

Pandey A. (2001). Common standards for genomics and proteomics. *Trends Genet* **17**: 696.

Porter DA, Krop IE, Nasser S, Sgroi D, Kaelin CM, Marks JR, *et al*. (2001). A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res* **61**: 5697–5702.

Proudfoot N. (1991). Poly(A) signals. *Cell* **64**: 671–674.

Pruitt KD, Maglott DR. (2001). RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* **29**: 137–140.

Riedy MC, Timm EA Jr, Stewart CC. (1995). Quantitative RT-PCR for measuring gene expression. *Biotechniques* **18**: 70–74, 76.

Riggins GJ, Strausberg RL. (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum Mol Genet* **10**: 663–667.

Rodriguez-Tome P, Lijnzaad P. (1997). The Radiation Hybrid Database. *Nucleic Acids Res* **25**: 81–84.

Salamov AA, Solovyev VV. (1997). Recognition of 3′-processing sites of human mRNA precursors. *Comput Appl Biosci* **13**: 23–28.

Schaefer C, Grouse L, Buetow K, Strausberg RL. (2001). A new cancer genome anatomy project web resource for the community. *Cancer J* **7**: 52–60.

Schena M, Shalon D, Davis RW, Brown PO. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467.

Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* **93**: 10614–10619.

Schwab M, Alitalo K, Klempnauer KH, Varmus HE, Bishop JM, Gilbert F, *et al*. (1983). Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. *Nature* **305**: 245–248.

Scott HS, Chrast R. (2001). Global transcript expression profiling by Serial Analysis of Gene Expression (SAGE). *Genet Eng* **23**: 201–219.

Sheets MD, Ogg SC, Wickens MP. (1990). Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res* **18**: 5799–5805.

Spieker N, van Sluis P, Beitsma M, Boon K, van Schaik BD, van Kampen AH, *et al*. (2001). The MEIS1 oncogene is highly expressed in neuroblastoma and amplified in cell line IMR32. *Genomics* **71**: 214–221.

Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. (2000). The cancer genome anatomy project: building an annotated gene index. *Trends Genet* **16**: 103–106.

Strausberg RL, Dahl CA, Klausner RD. (1997). New opportunities for uncovering the molecular basis of cancer. *Nature Genet* **15** (Special Issue): 415–416.

Strausberg RL, Feingold EA, Klausner RD, Collins FS. (1999). The mammalian gene collection. *Science* **286**: 455–457.

Ullman JD. (1988). *Principles of Database and Knowledge-Base Systems*. Computer Science Press: New York.

Velculescu VE, Vogelstein B, Kinzler KW. (2000). Analysing uncharted transcriptomes with SAGE. *Trends Genet* **16**: 423–425.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. (1995). Serial analysis of gene expression. *Science* **270**: 484–487.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al*. (2001). The sequence of the human genome. *Science* **291**: 1304–1351.

Yeang CH, Ramaswamy S, Tamayo P, Mukherjee S, Rifkin RM, Angelo M, *et al*. (2001). Molecular classification of multiple tumor types. *Bioinformatics*, **17** (Suppl. 1): S316–S322.

Zammatteo N, Jeanmart L, Hamels S, Courtois S, Louette P, Hevesi L, *et al*. (2000). Comparison between different strategies of covalent attachment of DNA to glass surfaces to build DNA microarrays. *Anal Biochem* **280**: 143–150.

Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, *et al*. (1997). Gene expression profiles in normal and cancer cells. *Science* **276**: 1268–1272.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 16

Proteomic Informatics

JÉRÔME WOJCIK and ALEXANDRE HAMBURGER

Hybrigenics
Paris, France

16.1 INTRODUCTION

16.1.1 A Definition of Proteomics

As genomics is the study of the set of genes in genomes, proteomics deals with the analysis of the 'proteome', that is the product of translation of the transcriptome.

The completion of the sequencing of bacterial and higher eukaryotic organisms marks the beginning of the post-genomic era. As more and more raw data become available, new challenges arise, namely handling these data and making sense out of them. Proteomics is a way of giving relevant meaning to these data by redefining them in a higher-level, function-oriented context, closer to what we may broadly call 'biological function'.

16.1.2 Challenge Compared to Genomics: Identification of 'Function'

The term 'proteomics' yields a new conception of the functional assignment issue in biology. 'Prote-' indicates that function is sustained by proteins, not by genes, and '-omics' proposes that function is defined 'in context'. The function of a protein is not solely an individual property of the protein but is defined as a combination of its biochemical interactions with its partners and the environment in which it exists. Information on the scale of the whole cell is therefore needed to comprehensively understand the function of proteins.

Protein sequence information is often an endpoint for the geneticist, for example, an amino acid substitution may be defined by a SNP. But as a matter of fact, this is just one element of many that can tell us about the properties of a protein. Other meaningful information can tell us a great deal more about the nature of proteins, such as 3D structure, post-translational modifications, half-life, phenotypic role, enzymatic activity or quantity (abundance). These properties have also been proven to be tissue- and subcellular localization-specific. Beyond the properties of the protein itself, protein interactions are a rather novel data form that have been shown to be amenable to high-throughput analysis (which will be discussed shortly). These methods are powerful tools to define proteins and pathways in context on the cellular scale. Ultimately this is the objective of genetics and hence proteomics is a critical step in the progression from candidate gene to validated disease gene.

With the completion of many genome sequences, including human, the aforementioned issue of finding a relevant context to study biological data in is even more acutely

felt. Many of the recent advances in proteomics have been made during the analysis of prokaryotic organisms. In this field more than any other, prokaryotes may point the way forward for analysis methods in higher eukaryotes, such as man, for these methods rely heavily on fully optimized and complete datasets, an ideal that we still struggle to achieve in studies of human material. We can safely assume that sequence data is not a sufficient and rich enough source of information to reach higher levels of understanding or meaningful definition of protein function. Indeed, a raw DNA sequence may be altered by several phenomena, making any assumption on function difficult. To name a few: alternative splicing may lead a single gene (or pre-mRNA) to produce many gene products (or mature mRNA) in eukaryotes. Further down the protein synthesis pathway, post-translational modifications may result in proteic cleavages, glycosylation, etc. The regulation of proteins is by itself an issue: post-transcriptional regulation of protein expression (changes in protein synthesis and degradation rates) induces no obvious correlation between protein and mRNA expression levels in humans (Anderson and Seilhamer, 1997) or in yeast (Gygi *et al*., 1999); time and space regulations may sometimes be partially uncovered by sequence analysis (proteic translocation between subcellular compartments may be linked to the presence of peptide signals which are cleaved when the protein reaches a mature state) but the subcellular localization *per se*, turnover, dynamic behaviour or lifetime of a protein cannot be directly linked to sequence analysis alone.

16.2 PROTEOMIC INFORMATICS

From the term 'Proteomic Informatics', we have already given an overview of what 'proteomics' may be. As for 'Informatics', Luscombe (2001) defines Bioinformatics as 'conceptualizing biology in terms of molecules (in the sense of physical-chemistry) and then applying informatics techniques (derived from disciplines such as applied mathematics, computer science and statistics) to understand and organize the information associated with these molecules, on a large-scale'. As high-throughput methods for biological data generation have been developed, we need powerful automated tools for analysing and understanding them. This is the goal of proteomic informatics. Data may be seen as a dense, fuzzy cloud of points in a complex, multidimensional space. It is the role of Bioinformatics to find a relevant subspace and project our data in a meaningful and understandable way that will enable us to reap the rewards of our data while not losing valuable information. At first glance, Proteomic Informatics may be seen only as a tool for data handling and visualization but its purpose is actually two-fold. On one hand, data may be displayed in a comprehensive way through the efficient use of bioinformatics tools and stored in rich databases that keep track of experimental settings. On the other hand, algorithms may be developed and improved to extract new information. As would befit bioinformatics tools aimed at proteomics applications, they should be able to process high quantities of data and conceptualize them as integral parts of a cellular context; hence the need to develop algorithms allowing reconstruction or inference of cellular pathways and protein–protein interaction maps.

16.3 EXPERIMENTAL WORKFLOW: CLASSICAL PROTEOMICS

The most frequently used high-throughput technology designed to study the proteome is aimed at identifying and quantifying the expression levels of proteins localized in specific protein complexes. This method is sometimes referred to as 'Classical Proteomics',

compared to 'Functional Proteomics' which concerns itself with the identification of interactions and cell processes. A typical approach consists in the separation of the various proteins of a cellular extract by gel electrophoresis followed by mass spectrometric analysis: comparison of the resulting experimental data with that available from sequence databases provides unique assignments for protein gel spots to their corresponding DNA sequences. Recent optimizations of the various steps provide one of the most powerful approaches in proteomics. The following section details the experimental workflow.

16.3.1 Proteome Purification

Sample preparation is the first and a crucial step in classical proteomics. The purer the sample, the more accurate the expression quantification and protein identification will be. Proteins can be extracted from whole cells (bacteria, yeasts...), tissues, or subcellular compartments (organelles). Purification methods include mainly centrifugation in density gradients, exclusion chromatography, affinity chromatography using for example peptide tags, antibodies (immuno-precipitation) or substrates (for reviews see Legrain *et al*. (2000) or Lee (2001)). A tandem affinity purification (TAP) involving a combination of two high-affinity tags linked to the protein of interest was also suggested as a general method for protein complex purification in mild conditions after expression in natural conditions (Rigaut *et al*., 1999) and was recently comprehensively applied to the yeast proteome (Gavin *et al*., 2002).

16.3.2 Proteome Separation: Electrophoresis

In the next step, the protein expression profile of the sample is typically deduced by 2D gel SDS-polyacrylamide gel electrophoresis (SDS-PAGE), a high-resolution technique for decomposing protein complexes of tenths of polypeptides (see Lee (2001) for review). Proteins are separated according to both isoelectric point (pI) and molecular weight (Mw), by a combination of isoelectric focusing and electrophoresis respectively. Spots are detected using colour stains, fluorescent dyes or radioactive labels (Figure 16.1).

Proteins can also be separated by classical 1D-PAGE but this requires reduction of the number of proteins in the cell extract, for instance by immuno-affinity purification (Ho *et al*., 2002) or TAP (Gavin *et al*., 2002).

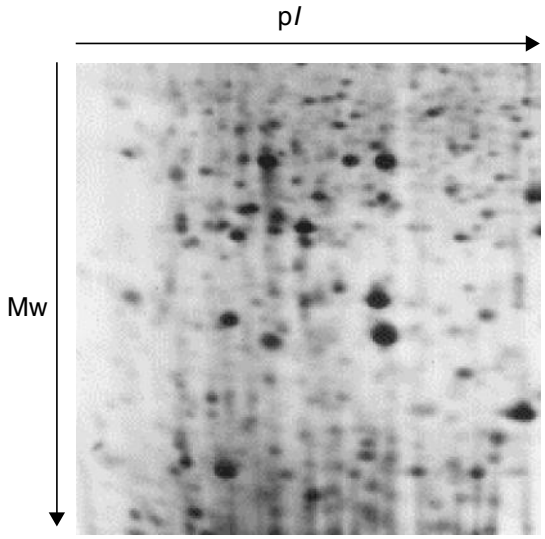


Figure 16.1 An example of 2D-PAGE. Proteins are identified by black spots after separation by electrical focusing (pI) and electrophoresis (Mw).

As SDS-PAGE becomes the most commonly used bidimensional protein separation method in proteomics, the technique is becoming standardized among different laboratories and databases of 2D gel images highlighting protein spots with appropriate links have been created for various proteomes (see Table 16.1).

16.3.3 Proteome Identification: Mass Spectrometry

Third, the separated protein spots on 2D gels are excised and digested in-gel with a protease (usually trypsin). The eluted peptides are then analysed by Mass Spectrometry (MS). Reaching a high level of sensitivity, automation and throughput for protein analysis, mass spectrometry has become one of the key technologies in the proteomics field.

Analysing femtomoles of protein materials is now routinely carried out using MALDI (Matrix-Assisted Laser Desorption/Ionization)/TOF (Time-Of-Flight)-based peptide mass fingerprinting, which provides a list of masses for the peptides contained in the digested 2D spot. Matching these against the list of calculated peptide masses from an appropriate protein sequence database characterizes the isolated protein (see for example, Houry *et al*., 1999).When the mass fingerprint is not found in databases, Tandem Mass Spectrometry (or MS/MS) can be used to sequence the polypeptides, thus providing sequence tags that could allow protein identification by sequence similarity screening of classical bioinformatics databases (for example EMBL by using BLAST (Altschul *et al*., 1997)). The combination of peptide mass fingerprinting followed by sequence tagging is a suite

TABLE 16.1 Main Online 2D-PAGE Proteomics Resources

Database	URL
Aarhus 2DPAGE database	biobase.dk/cgi-bin/celis
Aberdeen 2DPAGE	www.abdn.ac.uk/□mmb023/2dhome.htm
Argone protein mapping group	www.anl.gov/BIO/PMG/
Cyano2Dbase	www.kazusa.or.jp/cyano/cyano2D/
ES cell-2DPAGE	www.dur.ac.uk/□dbl0nh1/2DPAGE/
Harefield HSC 2DPAGE	www.harefield.nthames.nhs.uk/nhli/protein/
Maize Genome database	moulon.moulon.inra.fr
Maritime pine 2DPAGE	www.pierroton.inra.fr/genetics/2D/
Max-Planck Institut 2DPAGE	www.mpiib-berlin.mpg.de/2D-PAGE/
MDC Heart-2DPAGE	www.mdc-berlin.de/□emu/heart/
Parasite Host Cell Interaction 2DPAGE	www.gram.au.dk
Plant Plasma Membrane Database	sphinx.rug.ac.be:8080/ppmdb/index.html
SWISS-2DPAGE	www.expasy.ch/ch2d
SIENA-2DPAGE	www.bio-mol.unisi.it/2d/2d.html
SSI-2DPAGE	www.ssi.dk/en/forskning/tbimmun/tbhjemme.htm
TMIG 2DPAGE	proteome.tmig.or.jp/2D/
Université Paris 13 2DPAGE	www-smbh.univ-paris13.fr/lbtp/Biochemistry/biochimie/bque.htm
2DWGDB (WebGel)	www-lmmb.ncifcrf.gov/2dwgDB
WU Inner Ear database	oto.wustl.edu/thc/innerear2d.htm
Yeast 2DPAGE	yeast-2dpage.gmm.gu.se/
Yeast Protein Map (YPM)	www.ibgc.u-bordeaux2.fr/YPM/

of powerful techniques used to analyse and identify proteins (Quadroni and James, 1999; Yates, 1998).

One step further, MS coupled with High Performance Liquid Chromatography (HPLC) techniques and/or combined with biochemical techniques (immunoprecipitation) can provide shotgun identification of proteins in complex biological mixtures in order to study protein–protein interaction, to locate and identify single protein or protein complexes from a subcellular fraction. For instance, using a combination of HPLC and ESI (Electrospray Ionization)-MS, it has been shown that a large transmembrane protein (the lactose permease) could be analysed and studied quickly and with high accuracy (Whitelegge *et al*., 1999). High-throughput methods have also been designed to identify various post-translational modifications of proteins by mass spectrometry (Wilkins *et al*., 1999).

16.3.4 Building Protein Expression 'Networks'

Proteome-wide characterization allows the production of global maps of differentially expressed proteins. By comparing several sets of expression patterns under different conditions (for instance, wild-type versus mutant or normal versus diseased) or at different time stages, one can deduce clusters of co-regulated proteins that could be interpreted as a protein expression 'network'. Such differential protein expression networks have been applied for instance to the elucidation of cell pathways, the characterization of cell types or the identification of pathogenic agents (for review see Legrain *et al*., 2000). They are complementary to gene regulation networks produced by transcriptomics techniques (see Chapter 15).

Mass spectrometry also allows the identification of protein complexes, which could be conceptualized as clusters of the expression network. The technique was recently applied to detect yeast complexes on a proteome-wide scale (Gavin *et al*., 2002; Ho *et al*., 2002).

16.3.5 Analysing Protein Expression Data

Approaches to 2D gel image analysis may range from very basic to fairly complex. Several commercial 2D gel image analysis software packages are available that allow display, analysis and comparison of gel images, as well as determination, quantification and normalization of spots (Table 16.2). One can also use Flicker (Lemkin and Thornwall, 1999), a free web tool for comparing images from different internet sources. Given two gel images URL, Flicker loads the images and displays them in the web browser. They can be enhanced in various ways (spatial warping, pseudo 3-dimensional image sharpening…), while regions of interest can be 'landmarked' with several corresponding points in each gel image. One gel image is then warped to the geometry of the other and the two resulting images are compared visually in a third window (the 'flicker' window): as the two gels are rapidly alternated ('flickered'), the user can slide one gel past the other to visually align

TABLE 16.2 Some Gel Analysis Software

Software	Company	Reference
Melanie	Geneva Bioinformatics	www.expasy.ch/melanie
PDQuest	Bio-Rad	www.proteomeworks.bio-rad.com
Phoretix	Phoretix advanced	www.phoretix.com
Flicker		www.hi-beam.net

corresponding spots by matching local morphology. With such image analysis tools, an expert can locally visualize an expression network and formulate biological hypotheses. The next step is the automated numerization and database storage of protein expression patterns to allow high-throughput screening.

16.4 PROTEIN INTERACTION NETWORKS

If protein expression networks give information about co-regulation of proteins and their response to specific conditions, they are not completely informative about the biochemical function of gene products. Determining which other cell components interact with proteins addresses this issue. The function of a protein can be defined by the role it takes in cell pathways and the interactions in which it participates with other cell components (DNA, RNA, proteins, metabolites or lipids for instance). We distinguish here the interaction networks dealing only with proteins and produced by high-throughput experimental protocols from those containing heterogeneous factors (referred to as 'cell pathways'). The set of technologies used to produce interaction data on a large scale is referred to as 'Functional Proteomics'.

16.4.1 Experimental Technologies

Low-throughput technologies (co-immunoprecipitations, far-Western blots, 'pull-downs', etc, see Phizicky and Fields (1995) for review) are commonly used for studies on individual proteins. The study of interactions at the proteome level, however, requires high-throughput assays.

16.4.2 Yeast Two-Hybrid (Y2H)

The yeast two-hybrid system (Fields and Song, 1989) can detect interactions between two known proteins or polypeptides and can also search for unknown partners (prey) of a given protein (bait) (for review, see Vidal and Legrain, 1999). Yeast two-hybrid assay remains the main large-scale technology that is available to build protein interaction maps. Two strategies — namely the matrix approach and the fragment (or polypeptide) library screening approach — have been tested to find the most efficient way to explore proteomes for interactions (the interactome).

The matrix approach uses a collection of predefined open reading frames (ORFs), usually full-length proteins, as both bait and prey for interaction assays. Combinations of bait and prey can be assessed individually or after pooling cells expressing different bait or prey proteins. The intrinsic limitation of this strategy is that it tests only known proteins that are predefined. Y2H was first used to explore interactions among drosophila proteins involved in the control of cell cycle (Finley and Brent, 1994). Several studies have now been published for the yeast proteome, either comprehensive (Ito *et al*., 2000, 2001, Uetz *et al*., 2000) or using only a subset of specific baits (Newman *et al*., 2000).

The alternative Y2H assay strategy uses exhaustive libraries to screen for the identification of new protein interacting partners. Applying this library screening approach to functionally related proteins results in connection of uncharacterized proteins to specific pathways. It can be also applied to whole cellular interactomes. Screening numerous randomly generated fragments contained in the libraries also permits the determination of interacting domains defined experimentally as the common sequence shared by the selected overlapping prey fragments (Rain *et al*., 2001). This approach was first applied

to determine protein networks for the T7 phage proteome which contains 55 proteins (Bartel *et al*., 1996) and later applied to the yeast proteome focused on the RNA metabolism (Fromont-Racine *et al*., 1997) and to the human gastric pathogen *Helicobacter pylori* (Rain *et al*., 2001).

The two two-hybrid strategies are depicted in Figure 16.2. The pros and cons of each technology are discussed in a review (Legrain *et al*., 2001). Table 16.3 draws an inventory of major two-hybrid large-scale assays performed so far.

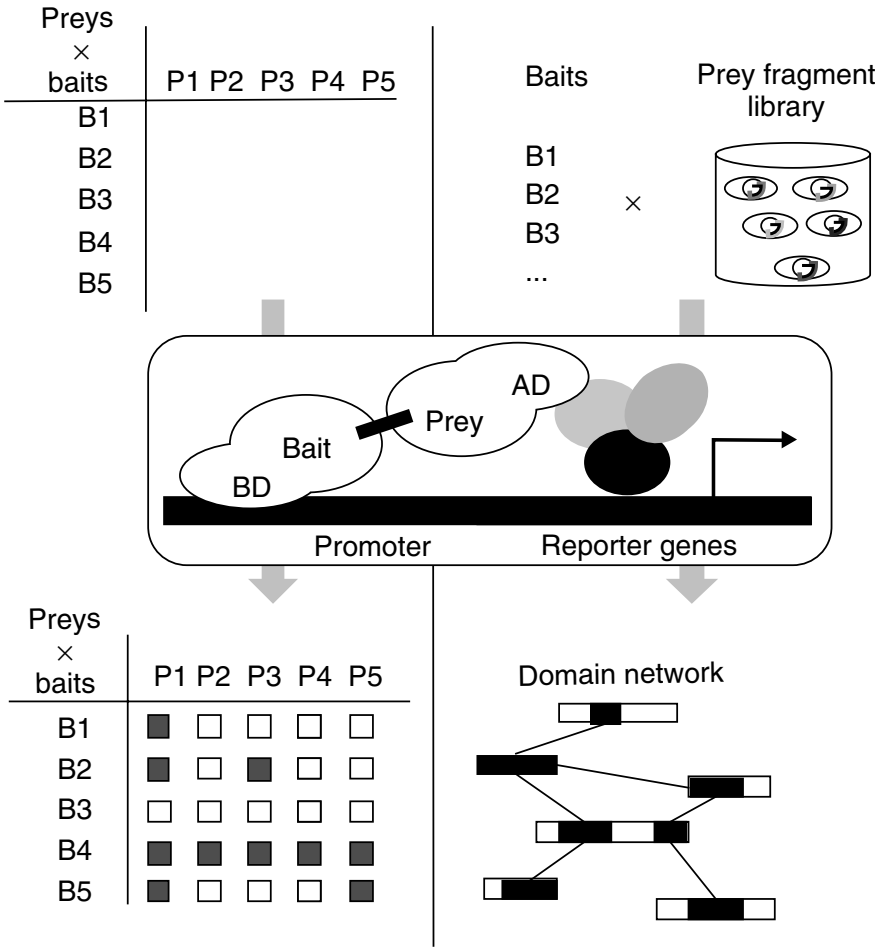


Figure 16.2 The yeast two-hybrid strategies. The central box schematizes the principle of the yeast two-hybrid assay: a protein domain that binds specifically to DNA sequences (BD) is fused to a polypeptide dubbed the 'bait' and a domain that recruits the transcription machinery (AD) is fused to a polypeptide dubbed the 'prey'. The basis of the assay is that transcription of a reporter gene will occur only if the bait and the prey polypeptides interact together. The matrix approach (first column) uses the same collection of proteins used as bait (B1–B5) and prey (P1–P5). The results can be drawn in a matrix where bait autoactivators (B4 for example) and 'sticky' prey proteins (P1 for example interacts with many proteins) are identified and discarded. The final result can be summarized as a list of interactions that can be heterodimers (B2–P3) or homodimers (B5–P5). The library screening approach identifies for each interacting prey protein the domain of interaction with a given bait. Sticky prey proteins are identified as fragments of proteins that are often selected regardless of the bait protein. An autoactivator bait can be used in the screening process with more stringent selective conditions.

TABLE 16.3 Key Figures in Large Scale Datasets for Protein–Protein Interaction Maps

Organism	Technology	Number of assays baits × preys	No. of interactions	Reference
Vaccinia virus	Protein array	Proteome × proteome	37	McCraith *et al.* (2000)
S. cerevisiae	Protein array	192 × proteome	281	Uetz *et al.* (2000)
	Pools of preys	Proteome × proteome	692	
S. cerevisiae	Pools of baits and preys	430 assays of pools (96 × 96)	175	Ito *et al.* (2000)
S. cerevisiae	Pools of baits and preys	3844 assays of pools (96 × 96)	841[a]	Ito *et al.* (2001)
S. cerevisiae	Protein array	162 × 162	213	Newman *et al.* (2000)
C. elegans	Protein array	29 × 29	8	Walhout *et al.* (2000)
	Library screening	27 × proteome	124	
HCV	Protein array	10 × proteome	0	Flajolet *et al.* (2000)
	Library screening	22 fragments × proteome	5	
S. cerevisiae	Library screening	15 × proteome	170	Fromont-Racine *et al.* (1997)
S. cerevisiae	Library screening	11 × proteome	113	Fromont-Racine *et al.* (2000)
H. pylori	Library screening	261 × proteome	1524	Rain *et al.* (2001)

[a] This number corresponds to highly significant interactions (more than three hits, see Ito *et al.*, 2001).

16.4.3 Other Technologies

Phage display technology is another assay used to screen a library of polypeptides for interaction with a target protein. Each polypeptide is expressed on the surface of a bacteriophage particle, as a fusion with a phage coat protein. This provides a physical link between the expressed polypeptide and its encoding gene. The phage-displayed polypeptide can be selected by binding to a target using affinity chromatography and further characterized by amplification and sequencing of the corresponding gene located within the phage particle. No protein–protein interaction map using phage display has been published so far either for an organism or an entire cell but the technology has a high-throughput potential (see for example Walter *et al.*, 2001). The technology is particularly suited for screening libraries of random polypeptide variants, such as antibody fragments

and can be combined to the complementary yeast two-hybrid technology in order to obtain more relevant results (Tong *et al.*, 2001).

Protein microarrays are also emerging in order to study protein–protein interactions. Known proteins are precisely spotted on glass substrates and used to probe interactions with peptides (Lueking *et al.*, 1999) or proteins (Haab *et al.*, 2001). A similar method was also tested to screen for small molecules (MacBeath and Schreiber, 2000).

16.5 BUILDING PROTEIN INTERACTION NETWORKS

16.5.1 From Experimental Results to Graphs

When the two protein partners are identified, a graph can be built where the vertices are the proteins (bait or prey) and the edges are the protein interactions. This step is trivial when the two partners are known beforehand, for example in the two-hybrid matrix approach, but requires post-processing when a partner is screened against a library and has selected a target/prey. In the latter case, the prey gene must be sequenced and identified in sequence databases using tools such as BLAST (Altschul *et al.*, 1997). When several experimental protocols are combined, for instance phage display and yeast two-hybrid (Tong *et al.*, 2001), one can decide whether to consider the totality of the interactions or only those common to both techniques, depending on the desired trade-off between false negatives and false positives (see below).

Moreover, in the two-hybrid strategy using fragment libraries, the functionally interacting domains can be precisely mapped on proteins: the common sequence shared by the selected overlapping prey fragments experimentally defines the smallest docking site selected by the bait (Rain *et al.*, 2001). The interaction network can then also be represented as a graph where the vertices are protein domains instead of full-length proteins.

16.6 FALSE NEGATIVES AND FALSE POSITIVES

One major drawback of the high-throughput experimental technologies described above is the generation of potential false negatives and false positives, depending on the assay conditions.

False-negative interactions are biological interactions that are missed because of incorrect folding, inadequate subcellular localization, lack of specific post-translational modifications etc. In yeast two-hybrid assays, the matrix approach is prone to generate a high level of false negatives (see Table 16.3), because only two assays are performed for each pair of proteins (bait versus prey, and reciprocally), whereas the fragment library approach allows testing of millions of potential interactions simultaneously. For instance, the two exhaustive studies of the yeast proteome (Ito *et al.*, 2001; Uetz *et al.*, 2000) have failed to recapitulate as much as 90% of interactions previously described in the literature (Ito *et al.*, 2001). The intrinsic limitations of the matrix approach concerning the choice of selective conditions can also explain this high rate of false negatives (for review see Legrain *et al.*, 2001).

Conversely, searching for many potential interactions, especially when screening a random fragment library, increases the chance of selecting biologically non-significant interacting polypeptides, thus leading to false positives. First, some bait proteins might have a predisposition to activate the transcription of reporter genes without specific interaction with any prey protein. These *auto-activator* bait proteins may randomly select

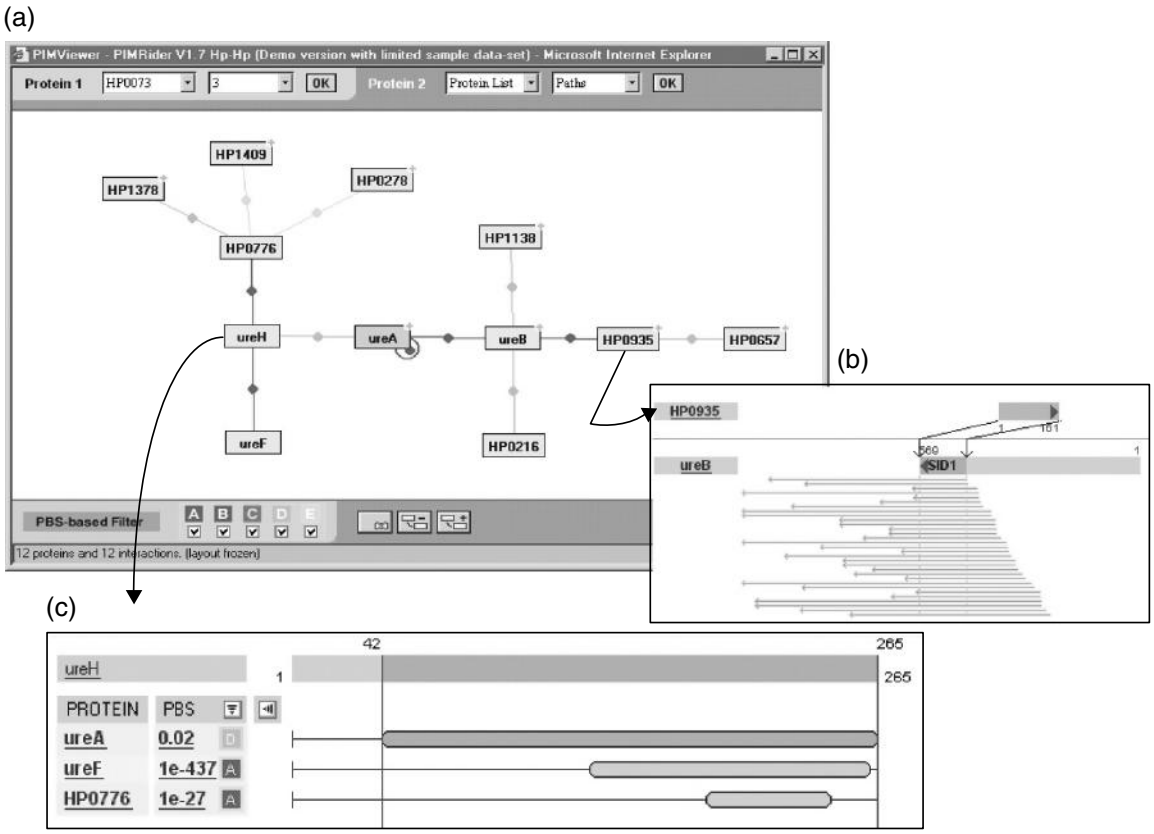


Figure 16.3 The PIMRider: an integrated exploration platform for protein interaction networks. The main window (a) displays the cell-wide protein interaction map as a graph and allows the biologist to navigate through the network, filter information depending on its reliability (PBS®) and focus on a particular pathway. Clicking on a specific interaction gives access to primary two-hybrid data (b) where interacting fragments and the computed Selected Interacting Domain (SID®) are positioned relative to the coding sequence of the two proteins. The biologist can make up his/her own mind about the interaction significance. All the interacting domains of one protein with its partners to formulate biological hypotheses, for instance about dominant negative interactors (c), can also be displayed.

prey proteins in addition to specific ones. Second, some chimeric prey proteins, dubbed *sticky* proteins, may similarly be non-specifically selected by many independent bait proteins. Discarding autoactivator bait proteins (that select many prey proteins) or sticky prey proteins (that are selected in many screens) leads to results with a reduced rate of false positives, although it may also mean a slightly increased number of false negatives (Ito *et al*., 2000, 2001). Less stringent filtering was used for the interaction network of *H. pylori*, based on a fragment library approach (Rain *et al*., 2001). A scoring scheme was designed that computes an E-value for each bait–prey interaction by comparing the observed pattern of selected prey fragments with the theoretical pattern that would be obtained by randomly picking fragments in the library. At the end, each interaction yields a reliability value (see Figure 16.3).

16.7 ANALYSING INTERACTION NETWORKS

The first protein interaction databases available on the internet provided a basic display of the alphabetical protein interaction list. An interaction is represented by its two protein

TABLE 16.4 Main Protein–Protein Interaction Databases

Database	URL	Reference
EcoCyc	ecocyc.org/ecocyc/ecocyc.html	Karp *et al.* (2000)
BIND	www.bind.ca	Bader and Hogue (2000)
Cellzome	yeast.cellzome.com	Gavin *et al.* (2002)
CuraGen portal	portal.curagen.com	Uetz *et al.* (2000)
DIP	dip.doe-mbi.ucla.edu	Xenarios *et al.* (2000)
FlyNets	gifts.univ-mrs.fr/FlyNets/	Sanchez *et al.* (1999)
Interact	bioinf.man.ac.uk/interactso.htm	Eilbeck *et al.* (1999)
MIPS	www.mips.biochem.mpg.de	Mewes *et al.* (2000)
PIM Rider	pim.hybrigenics.fr	Rain *et al.* (2001)
ProNet	pronet.doubletwist.com	

partners, sometimes with basic annotations or cross-references to other protein databases. Some websites also propose packages to graphically display interaction networks (Mrowka, 2001). The main protein–protein interaction sources are listed in Table 16.4.

However, a simple list of interactions poorly tackles the issue of result reproducibility. To evaluate false positives and reproducibility, access to primary data is necessary. For example, the interactions listed at the MIPS (Mewes *et al.*, 2000) only present a brief indication of the experimental source, such as 'two-hybrid' or 'co-immunoprecipitation', without any quality clue or reference to the source experiment or laboratory. Bioinformatics tools are now emerging to tackle this issue, such as the PIM Rider® (Rain *et al.*, 2001) which gives access to primary data (see Figure 16.3b).

Visualization software is in parallel being enriched with options to help the biologist in his/her discovery process. They let the user search for interaction paths between two given proteins, filter displayed interactions depending on their reliability value or simultaneously display all interacting domains identified in one specific protein (see Table 16.4 for examples, such as PIM Rider from Hybrigenics (Rain *et al.*, 2001), PIScout from LION Biosciences, or the visualisation tool of DIP (Xenarios *et al.*, 2000)).

16.8 CELL PATHWAYS

Cell pathways extend protein interaction networks by integrating interactions with lipids, small molecules (e.g. metabolites), RNA, DNA etc. They are mainly deduced from a compilation of literature resources, contrary to protein interaction networks that are technology-driven results.

16.8.1 Metabolic Pathways

The metabolism of living systems and their evolution have been investigated for a long time. The fluxes of metabolites inside a cell and the cascades of enzymatic reactions leading from one compound to another have been depicted in charts, that is, heterogeneous interaction networks mixing small molecules (metabolites) and proteins (enzymes). For example, Figure 16.4 illustrates the pyruvate metabolic pathway: the circles represent the small molecules that are the vertices of the metabolic network, whereas edges are catalytic reactions and are labelled with boxed enzymes. Several databases regroup information about these cell networks, especially for prokaryotic organisms (Kanehisa and Goto, 2000;

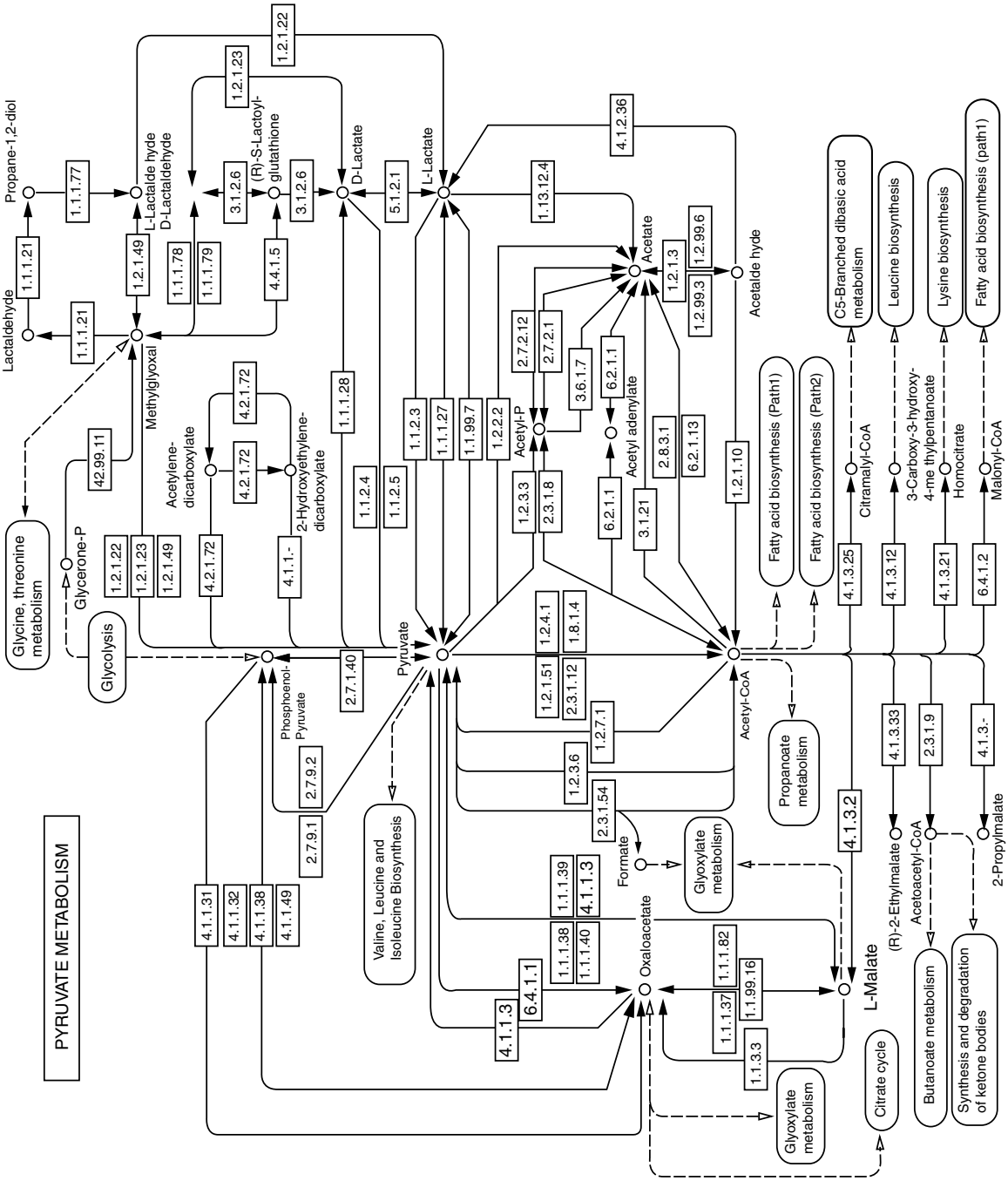


Figure 16.4 The pyruvate metabolism pathway. Reference 00620 taken from the *Kyoto Encyclopaedia of Genes and Genomes* (Kanehisa and Goto, 2000). Small circles represent molecule compounds and square boxes represent enzyme proteins, referenced by their EC numbers.

Karp *et al*., 2000; Selkov *et al*., 1998). Enzymes are referenced by their EC (Enzyme Commission) number, a system which overlays a functional hierarchy on enzymes (see Bairoch (2000) for a review).

16.8.2 Signal Transduction Networks

The signal transduction pathways are particular instances of internal cell pathways. They describe the cascades of molecular interactions from the reception of an extracellular signal (e.g. binding of a cytokine to its receptor) to the activation of transcription factors triggering the transcription of specific genes. The signal transduction networks are generally described in terms of physical interactions between proteins (e.g. binding or phosphorylation, etc; see Figure 16.5).

16.8.3 Gene Regulation Networks

Downstream of the signal transduction pathways a complex array of gene regulation networks takes place. The transcriptional regulatory networks mix heterogeneous physical interactions (protein–protein, protein–DNA, and protein–RNA) and genetic interactions (activation, inhibition, etc). Gene regulation networks are however still more studied at a higher level of abstraction (see Chapters 13 and 15).

Signal transduction and regulatory pathways have been constructed from individual experiments and stored in dedicated databases such as, SPAD http://www.grt.kyushu-u.ac.jp/spad/, TRANSFAC (Heinemeyer *et al*., 1999), or MIPS (Mewes *et al*., 2000).

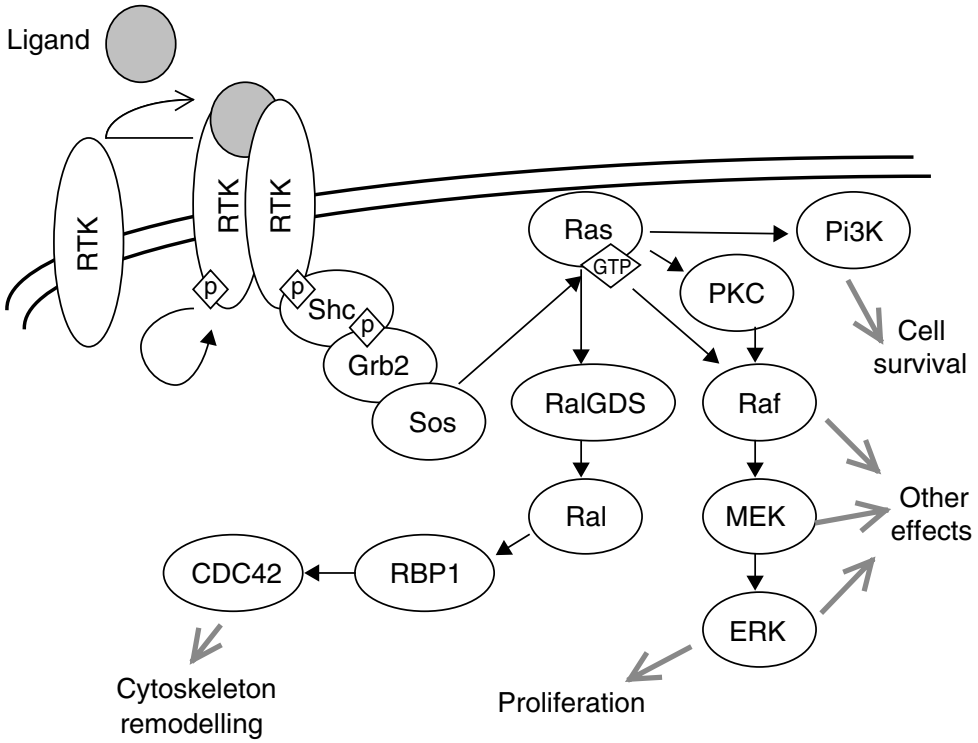


Figure 16.5 Signal transduction networks of TK receptors. The binding of a ligand to its tyrosine kinase receptor (RTK) provokes the dimerization of the receptor and the initialization of several intra-molecular signalling cascades, involving physical interactions and activation (black arrows: phosphorylation (P), GTP-binding (GTP), and others). One signal pathway triggers several biological effects (grey arrows).

These databases have allowed researchers to computationally predict regulatory networks, for example, Pilpel *et al*. (2001), computationally predicted an extensive transcriptional regulatory network in yeast by combinatorial analysis of promoter elements.

16.9 PREDICTION OF PROTEIN NETWORKS

16.9.1 Prediction of Functional Networks by Comparative Genomics

With the completion of many genome sequences, new techniques are emerging to predict the function of gene products by analysing the genes on a genome scale and comparing genomes between organisms. This new set of methods dubbed 'comparative genomics' has allowed the prediction of *functional* links between many proteins (see Eisenberg *et al*. (2000) for review).

Comparing genomes means comparing sequences of genes and establishing similarity links between genes means identifying orthologues, i.e. genes sharing the same function across organisms. In the following prediction method, the identification of orthology is often reduced to the detection of a significant sequence similarity, that is below a fixed E-value threshold, in a sequence similarity search (such as BLAST). Implications of this statement on prediction accuracy will be discussed below.

16.9.2 Gene Fusion Events

The gene fusion event method was first introduced by Marcotte *et al*. (1999a) and extended thereafter by other works (Enright *et al*., 1999; Marcotte *et al*., 1999b). The method is based on evolutionary interaction hypotheses. Basically, if two genes A and B participate in the same function, they are likely to be fused together during evolution to enhance the effective concentration of the fused gene product. Few mutations can then appear between the proteic domains from A and B. If genes A and B are once again separated, their products could still physically interact (Figure 16.6). Thus, if two separate genes in a given organism are fused together in another organism, they are likely to be functionally linked, that is to participate in the same structural complex, in the same biological pathway, in the same biological process or sometimes to physically interact (see examples in Figure 16.7). However, one cannot distinguish between these four kinds of functional links without extra information. The gene fusion event method is often referred to as the *Rosetta-stone* method (Marcotte *et al*., 1999a) in reference to the Rosetta stone which allowed Champollion to make sense of hieroglyphs ('word fusion') by comparing them to Greek and Demotic (languages using 'unitary' words).

The gene fusion event method was applied to the prediction of the protein functional network of *Escherichia coli* by comparing its genome to a set of 22 genomes of archaeal, bacterial and eukaryotic species (Tsoka and Ouzounis, 2000). In terms of participation in fusion events, a three-fold preference was evidenced for metabolic enzymes compared with control sets. It is worth mentioning that 76% of the detected pairs of enzymes participating in fusion events are known to be subunits of an enzymatic complex in the EcoCyc database (Karp *et al*., 2000; Table 16.4). The fusion event method thus seems to be able to detect physical interactions for metabolic enzymes.

16.9.3 Gene Neighbourhood

It was postulated for a long time that the way genes are organized in clusters in bacterial chromosomes is probably the result of an evolutionary constraint. The completion

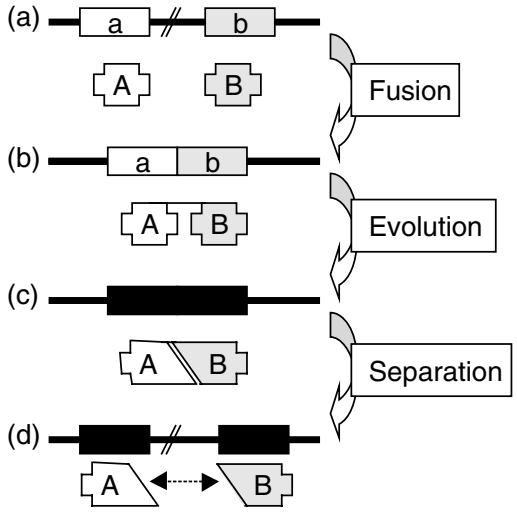


Figure 16.6 Underlying hypotheses of the gene fusion event method. This figure represents a model for the evolution of protein–protein interactions. If two genes a and b, originally separated in the genome (a), are fused together during evolution (b), the resulting chimeric protein A-B could mutate to develop intra-molecular contacts between A and B domains (c). Then, if the two initial genes are once again separated in genomes, the corresponding gene products A and B could still physically interact, or at least be functionally linked (d).

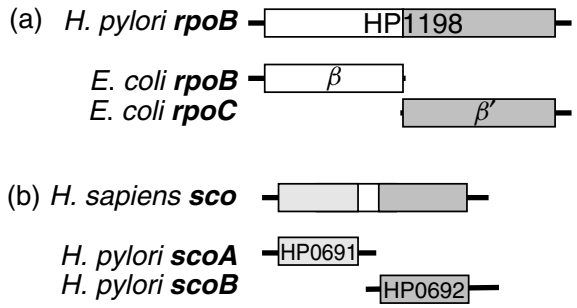


Figure 16.7 Examples of gene fusion events. (a) The β and β' subunits of the DNA-dependent RNA polymerase are encoded by two separate genes in most eubacteria and archea, but are fused together in a single gene in *Helicobacter pylori* (HP1198). These two subunits are known to be part of the RNA polymerase holoenzyme complex. (b) Similarly, the α and β subunits of the succinyl-CoA transferase in *H. pylori* (HP0691 and HP0692, respectively) are fused together in human and the corresponding gene products are predicted to physically interact in two-hybrid screens (Rain *et al*., 2001).

of many genome sequences now allows testing of this hypothesis at a comprehensive level. Dandekar and co-workers first analysed three triplets of sequenced genomes to identify conserved gene pairs (Dandekar *et al*., 1998). About 100 genes were found to be conserved as pairs, among them 75% of the encoded protein pairs physically interact. This suggests that conservation of gene order and physical interaction of encoded proteins are evolutionarily correlated.

Overbeek *et al*. (1999) extended this kind of analysis by building synteny groups, i.e. gene clusters across organisms, in order to infer functional links. They defined a gene cluster as a set of genes located on the same strand, and in which the maximal

intergenic distance is 300 base pairs. If two genes X_A and Y_A in a given cluster of genome A have orthologues X_B and Y_B in a cluster of genome B, they are defined as functionally coupled. A coupling score is also derived depending on the number of organisms in which orthologous pairs are found and the phylogenetic distances between these organisms and A.

The use of this gene neighbourhood method is obviously more efficient for microbial genomes with their conserved gene organization. But it may also be extended for eukaryotes where operon-like cluster structures have been observed (Wu and Maniatis, 1999).

16.9.4 Phylogenetic Profiles

A phylogenetic profile is defined as the occurrence pattern of orthologues for a given gene in a set of reference genomes (Pellegrini *et al*., 1999). It describes the absence or presence of a particular gene across this set of genomes (Figure 16.8). If two proteins have the same phylogenetic profile across these genomes (for instance P1 and P2, as well as P4 and P6 in Figure 16.8), it is assumed that they are functionally linked because they have probably co-evolved.

The major underlying hypothesis of the method is that orthologues, that is proteins having exactly the same function, are correctly identified. Moreover, all the reference genomes must be completely sequenced to avoid false-negative information. Note also that paradoxically if the identification of orthology heavily relies on sequence similarity, the phylogenetic profile method is referred to as a sequence-independent clustering algorithm, since proteins that are functionally linked in this way, i.e. that have the same phylogenetic profile, do not share sequence similarity in general.

16.9.5 Combination of Several Methods

Each of the previously described methods predicts functional links between proteins according to evolutionary and sequence-based hypotheses. Combining these approaches theoretically minimizes the false-positive prediction rate. Eisenberg and colleagues combined five types of protein–protein interaction links to build a functional linkage network for yeast, three of them are predictions from bioinformatics algorithms, two others are derived from experimental data (Marcotte *et al*., 1999b):

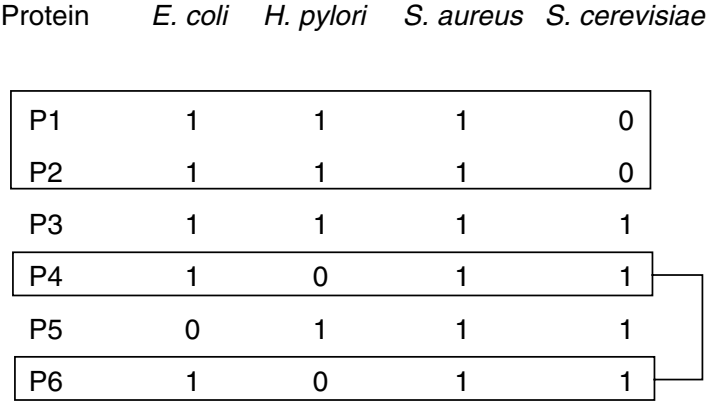


Figure 16.8 Clustering by phylogenetic profiles. The presence or absence of six proteins labelled P1 to P6 is indicated by 1 or 0, respectively, in four genomes. Proteins with the same profiles are boxed.

- Links from the Rosetta-stone method
- Links from the phylogenetic profile method
- Links between yeast proteins that have *Escherichia coli* homologues linked in metabolic pathways, as defined in the EcoCyc database (Karp *et al*., 2000)
- Links from known physical interactions in the DIP database (Xenarios *et al*., 2000)
- Links between proteins whose mRNA levels are correlated in cell cycle microarray experiments (Spellman *et al*., 1998)

The combination of these five networks represents over 93,000 pair-wise links between yeast proteins (about 30 links per protein), indicating a potentially high proportion of false positives. However, taking into account only 'highest confidence links', defined as links found by any two out of the three prediction methods or deduced from one of the two experimental techniques, reduces the number of links to 4130 (about 5%).

16.9.6 Inferences Across Organisms

Once a protein network is built for a given organism (by experimental or predictive methods) one might wonder how to transport it to other organisms. The classical inference mechanism involves two major steps:

1. A correspondence is established between proteomes, classically by identifying orthologues between organisms by sequence comparison.
2. The interaction links in the source protein network are transported to the target proteome along this correspondence.

The accuracy of these inference processes is highly dependent on the criteria chosen for orthology (i.e. conservation of function). Caveats of inferences will be further discussed in Section 16.10.

16.9.7 Protein Interaction Inferences

The inference process can be applied to all types of protein networks. It was recently tested on protein interaction methods (Wojcik and Schächter, 2001). An inference method similar to the one described above (correspondence according to sequence similarity on full-length sequences), referred to as the 'naive' method was assayed together with another method, dubbed the 'Interacting Domain Profile Pair' (IDPP) method, that combines sequence similarity searches with clustering based on interaction patterns and interaction domain information.

The principle of the IDPP method is illustrated by the prediction of a protein interaction network for *E. coli* from an experimental protein interaction map for *H. pylori* (Rain *et al*., 2001) in Figure 16.9. From the 1524 interactions in the original *H. pylori* network, the IDPP method led to 881 interaction predictions, connecting 412 proteins of *E. coli* (9.6%). Compared to the naive method, the IDPP method yields 35 additional, highly domain-specific, predicted interactions. The use of sequence similarity searches restricted to interacting domains rather than full-length proteins increases the sensitivity of the method. Similarly, the use of interacting domain clusters instead of single interacting domain sequences allowed the detection of homologies at lower levels of sequence similarity (see Figure 16.10 for an example). Six-hundred and fifty-one interactions were predicted by the naive method but not by the IDPP method. Two hundred and fifty-two

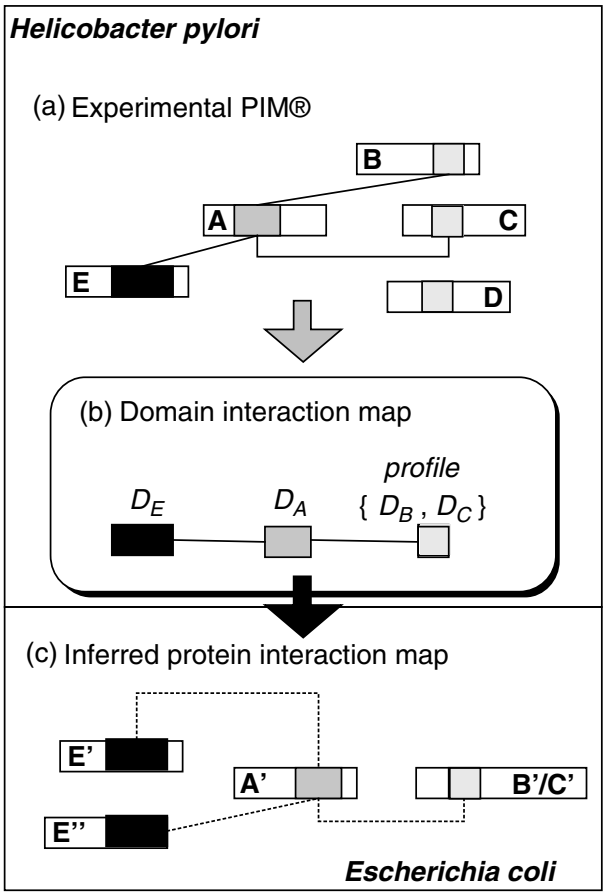


Figure 16.9 The interacting domain profile pair method. From the initial protein interaction map of *H. pylori* (a), an abstract domain cluster interaction map is derived (b). Domains are clustered together if (i) they share a significant sequence similarity and (ii) they share a common interaction property with a third partner (e.g. interacting domains of proteins B and C both interact with A). Each domain or profile of domains is then used as a probe to screen a library of *E. coli* protein sequences and domain cluster interactions are transferred (c).

of these 651 interactions were demonstrated to be false positives using the naive method since the prediction is achieved through sequence similarity of a region that does not contain the interacting domain. The 399 remaining interactions were obtained through sequence similarity that was significant when considering the whole protein but not when considering the shorter interacting domain and thus, might be considered as potential false positives.

16.10 ASSESSMENT AND VALIDATION OF PREDICTIONS

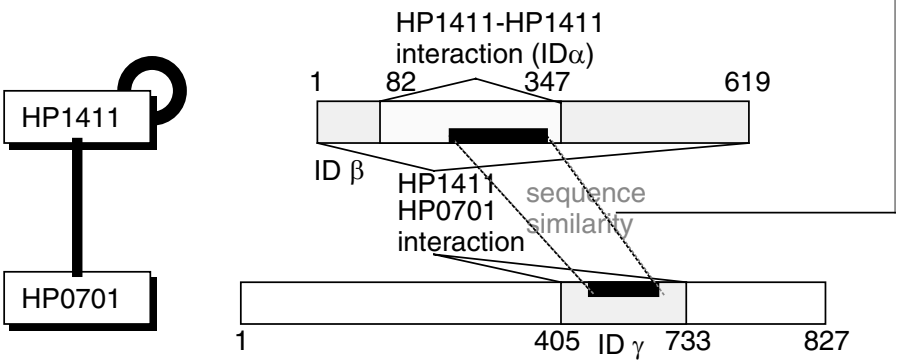
The methods described above predict protein networks. Each prediction method is based on a specific biological hypothesis and yields a set of given parameters both of which must be validated. The *validation* of bioinformatics predictions means the comparison of predicted results with the *state of the art* of biology. We distinguish here automated validation methods, that are systematic, reproducible, comparable and easy to perform but often yield weak biological confirmation, and manual validation methods, that are much more biologically informative but also more biased and laborious. We do not discuss here the validation of prediction methods *per se* but only the validation of predicted results.

(a)

```
HP1411 197 VKDFMSSKEFVAKYRYTPKQNTERAKKLQSYLENKRDFIGFVQALNSLKDNPQDPFLPNEETSFLVFAN
           ..:..... .:... :  :  ..:.  :..: ...:   : ... .. .:.  . :.  .  ..:.: .
HP0701 498 IEDLIANEPMVVSMSY--KGYVKRVD-LKAYEKQNRGGKGKLSG-STYEDDFIENFFVANTHDILLFIT

HP1411     EPTIVFNLRDYLLVLA-QIFNQQAICYCESKCPIELINAS-PGKDFNKTQDSFPDIKFSTPNQLEQSLN 332
           .   ...:. : .  : .:   .::    :  : : : :.   :::.  ..    . : : : . .  :
HP0701     NKGQLYHLKVYKIPEASRIAMGKAIVNLISLAPDEKIMATLSTKDFSDERS----LAFFTKNGVVKRTN 627
```

(b)



(c)

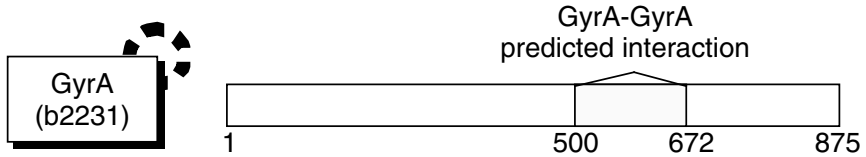


Figure 16.10 Prediction of GyrA homodimerization in *E. coli* by the IDPP method. In the *H. pylori* reference protein interaction map, the β interacting domain (ID) of HP1411 interacts with ID γ of HP0701 and HP1411 interacts with itself through ID α (b). When the IDPP method is applied, ID α and ID γ are clustered together since they both interact with the same region of HP1411 (b) and they share a sequence similarity (region 197–332 of HP1411 and region 498–627 of HP0701, 103 amino acid overlap, 32% of identity, (a)). This leads to the creation of a 'homodimer' profile pair connecting the α/γ domain profile with itself. When used as a probe to screen an *E. coli* protein sequence library, the α/γ domain profile selected a 172-amino acid-long domain on the GyrA protein, and GyrA was predicted to interact with itself through this domain (c). This prediction is confirmed by the literature: GyrA is known to form an A2–B2 complex with GyrB.

16.10.1 Automated Validations

The most widely used validation method is the 'keyword retrieval' technique. The principle is simple: if two proteins are linked together in the protein network, one compares their keywords according to a specific biological annotation and if they share similar keywords, the weight of the link is reinforced. The percentage of shared keywords at the network level is compared to a theoretical background noise to evaluate the global validity of the prediction. For instance, the keywords can be SWISS-PROT annotation keywords or functional categories (Jenssen *et al*., 2001; Marcotte *et al*., 1999b; Wojcik and Schächter, 2001). However, this validation method relies heavily on database annotations that are always reductive and sometimes false. For example, Marcotte *et al*. (1999a) noted that 'even truly related proteins show only a partial SWISS-PROT keyword overlap'. In this case they observed only a 35% overlap. Thus, this method, while significantly better than

random noise, probably gives a poor biological validation. Cross-validating protein interaction predictions by comparing annotations of both partners is also very dependent on the existence, the format and the quality of these annotations.

The second idea is to consider that a prediction, here a functional link between proteins, made by several independent methods is more reliable (in fact the random background noise of co-occurring independent facts is lower). This was used for instance to define high-confidence links in protein networks (Marcotte *et al*., 1999b) or to assess interaction predictions against physical location of genes in prokaryotic genomes (Wojcik and Schächter, 2001). In that case, there is a caveat to assess the real independence of prediction methods since the majority of them are sequence based. Basically, the more independent the prediction methods, the more relevant (in terms of false positives) the overlapping results will be.

Finally, predicted protein–protein links can be evaluated by checking their existence in dedicated databases, such as MIPS (Mewes *et al*., 2000), DIP (Xenarios *et al*., 2000) or OMIM (Hamosh *et al*., 2000). For instance these databases can be used to validate networks predicted from literature mining (Jenssen *et al*., 2001). These manually curated databases however regroup heterogeneous information and one must be cautious about data source quality. Predictions can also be compared to other types of data, such as gene clusters deduced from microarray data (Jenssen *et al*., 2001). In both cases, the significance of the predictions is evaluated by calculating the fold improvement over a virtual random experiment and/or the correlation between the two datasets.

16.10.2 Manual Validations

Using manual validation, each predicted interaction link between two proteins of a network is assessed by manually comparing the annotations in public databases, by checking literature references of each protein partner. This method is obviously low-throughput and by essence biased, but can lead to interesting conclusions about protein network quality.

It was for the first time applied to the assessment of inferred protein interactions from *H. pylori* to *E. coli* (J. Wojcik *et al*., unpublished data). The inference process is based on clustering and a definition of orthology restricted to the interacting protein domains (Wojcik and Schächter, 2001). The true positive prediction rate was evaluated to be at least 12%, i.e. at least 12% of the 1280 predicted interactions make biological sense according to biological curators. Three main causes were identified to explain predictions that are not confirmed by the literature: (i) predictions are true positives but are not yet referenced in the literature; (ii) one of the protein functions in the source interaction was completely lost during evolution (the corresponding gene has only paralogues in *E. coli*); or (iii) the source interaction is a false-positive result. The comparison of these exact but not statistically significant results with those obtained by automated validation by keyword retrieval (Wojcik and Schächter, 2001) emphasizes the need to have real and exhaustive reference datasets in order to validate predictions.

16.10.3 Literature Mining

The literature mining method, sometimes called ‘Information Retrieval’, can be viewed both as an assessment method to predict protein networks and as a prediction method *per se*. Assuming that the major part of current biology knowledge is contained in scientific literature, the parsing of titles, headings, abstracts and/or full texts of articles should enable us to extract links between genes or proteins and then build networks. Several techniques exist to perform this parsing, including linguistic methods that tag parts of words (e.g.

Ono *et al*., 2001) or statistical methods that estimate discriminating word distributions (e.g. Marcotte *et al*., 2001). One major issue in these studies is the establishment of an unambiguous nomenclature for gene or gene product names. Gene name dictionaries can be created from various nomenclature databases such as HUGO, LocusLink or OMIM, but problems remain due to insufficient synonym definition, synonym variations and gene families with fuzzy naming conventions.

A recent work aimed to analyse over 10 million MEDLINE records to detect and count human gene symbols or names co-occurring in titles or abstract. This resulted in a protein interaction network containing about 140,000 interactions connecting 7512 human genes (Jenssen *et al*., 2001). This is the largest protein network predicted from literature mining so far. For now mining literature is more profitably used to help the scientist by screening abstracts and reducing the number of articles to read. This is used to enrich the Database of Interacting Proteins (DIP) (Marcotte *et al*., 2001).

16.11 EXPLOITING PROTEIN NETWORKS

Once a protein network is experimentally built or predicted or inferred by bioinformatics algorithms, it represents a valuable source of information to understand molecular mechanisms on the scale of a whole cell, either by assigning function to gene products in context (local analysis of the protein map) or by analysing the global network shape and suggesting biological hypotheses.

16.11.1 Functional Assignments: the 'Guilt-By-Association' Rule

The first attempts to assign function used 'guilt-by-association' methods to annotate proteins on the basis of the annotations of their interacting partners or, more generally, of the proteins sharing a common property in a given cluster (Mayer and Hieter, 2000).

For example, a set of yeast protein interactions described in the literature or revealed by large-scale two-hybrid screens was analysed through a clustering method (Schwikowski *et al*., 2000) based on cellular role and subcellular localization annotations from the Yeast Proteome Database (Costanzo *et al*., 2000). The function of an uncharacterized protein is assigned on the basis of the known functions of its interacting partners. A function was assigned to 29 proteins (out of 554) that have two or more interacting proteins with at least one common function.

However, 'guilt-by-association' functional assignments must be used with caution. First the predictions are highly dependent on the database function annotations which are often reductive (only one keyword) and sometimes false. Poorly defined annotations can gather different concepts and induce biologically non-significant clustering. The assignments also obviously depend on the quality of the source protein network. If there are too few connections or, on the contrary, if there are too many false-positive connections to a protein node, the guilt-by-association would lead to erroneous conclusions. This point is especially crucial with two-hybrid interaction data, for which false positives represent highly connected nodes in the network.

Last, but not least, a major hurdle in this kind of automated function annotation method, common to all bioinformatics prediction algorithms, is the absence of an independent reference dataset and validation methods. For instance, the 29-function assignments made in the former study were compared with the corresponding high confidence links obtained in the study of Marcotte *et al*. (1999b) which were themselves partially predicted from interactions listed at MIPS, one of the yeast protein interaction databases used in the original study (Schwikowski *et al*., 2000). This exemplifies the fact that predictions must

be used with caution: the oversight of the initial study hypothesis and the deficiency in independent data sources could lead to biased conclusions.

Bioinformatics clustering of protein interactions still represents a powerful annotation tool which will become more and more useful as the interaction data accumulate and their quality improves. However, in order to be used successfully for appropriate functional annotation, the data needs to be stored in elaborate structures that allow each individual scientist to test their own hypothesis against complex heterogeneous primary data and then to design further experiments to validate the functional assignment.

16.12 DEDUCING PREDICTION RULES FROM NETWORKS

Given a protein interaction network and assuming that it is complete enough and has a low rate of false positives, one can deduce from the list of protein–protein interactions some biological information at a molecular level. We give here two examples of statements deduced from the analysis of comprehensive interaction maps that could *a posteriori* be used to predict protein interactions.

16.12.1 Domain–Domain Interactions

Two independent groups have analysed the available protein–protein interaction network of *Saccharomyces cerevisiae* in terms of domain–domain interactions (Ito *et al*., 2000, 2001; Mewes *et al*., 2000; Uetz *et al*., 2000; Xenarios *et al*., 2000). The first group (Park *et al*., 2001) considered protein structural domains from the SCOP classification (scop.mrc-lmb.cam.ac.uk/scop/; Murzin *et al*., 1995) and the second group (Sprinzak and Margalit, 2001) studied motifs from the InterPro database (www.ebi.ac.uk/interpro/; Apweiler *et al*., 2001). The basic idea is to count the co-occurrence of pairs of domains in interacting proteins, and to compare it to a theoretical background, in order to use over-represented domain pairs as predictors.

16.12.2 Correlated Mutations

As stated previously, the interactions in which a protein participates define its function. The specificity of these interactions is essential for the protein function to some extent. Thus, if the protein evolves and some point mutations occur at the interaction interface, 'complementary' mutations should also occur on protein partners to guarantee the interaction specificity. This hypothesis was developed by Pazos *et al*. (1997) who showed that correlated mutations in interacting domain pairs occur favourably close to the structural protein–protein interface. They proposed the use of this information to help to discriminate between several docking propositions when the 3D structure of both protein partners is known.

16.12.3 Analysis of the Shape of Protein Networks

Rather than focusing on a specific protein node in a protein network, one can analyse the whole interaction map to deduce biological hypotheses on the cellular scale. Jeong and co-workers published such an analysis of the public yeast protein interaction map (Jeong *et al*., 2001). They showed that this network forms a scale-free network: the probability that a given protein interacts with k partners follows a power law. This kind of structure is particularly tolerant to random attacks on one hand, and fragile against attacks

targeted on the most connected nodes on the other hand (Albert *et al*., 2000). Similar non-homogeneous network structures were also evidenced for metabolic networks (Jeong *et al*., 2000) and another protein interaction map in bacteria (Rain *et al*., 2001). The authors established a positive correlation between connectivity and lethality: highly connected proteins are three times more likely to be essential, i.e. the yeast cell dies if the corresponding gene is deleted. This correlation has been attributed to evolutionary selection.

Although the existence of such a correlation makes biological sense, one should probably wonder about the relative weight of technological bias in establishing it. Jeong's work indeed rests mainly on interaction data produced by one systematic two-hybrid system in yeast. The technology is prone to induce false negatives and false positives, as illustrated and commented on in a more recent similar study (Ito *et al*., 2001). The corresponding protein interaction network which contains 1870 proteins (31% of the whole yeast proteome), is not complete. Its shape would probably be different if all 'real' interactions were known. Proteins that exhibit few interacting partners in this network could actually represent highly connected nodes. Conversely, false positives in the two-hybrid system are likely to result in highly-connected nodes of the network: so-called 'sticky prey' proteins bind 'by chance' to many independent bait proteins. The correlation between lethality and centrality in networks evidenced by Jeong and co-workers, could actually be much stronger if genes that are both non-essential and highly connected on the one hand and genes that are both essential and poorly connected on the other hand, proved to be the consequences of a technological bias in data.

16.12.4 Precautions for Protein Networks

To conclude, both local 'guilt-by-association' functional assignment rules and global network analysis methods are fragile against poor interaction data quality or incompleteness. They can thus hardly produce reliable 'local' conclusions and the fact that the conclusion appears biologically meaningful is not evidence of the validity of the demonstration *per se*.

It is, moreover, imperative to assess the technological data bias prior to analysing networks and formulating biological conclusions. Ideally, the false-negative rate should be minimized by building comprehensive networks and false positives should be filtered out by independent bioinformatics or experimental validations. Meanwhile, both should be assessed using technology-specific reliability score assignments (Rain *et al*., 2001).

16.13 CONCLUSION

Since proteins and RNA sustain function rather than genes, and since function can no longer be considered as an individual property of each molecular actor taken independently from others, proteomics has appeared as the post-genomic method of choice. High-throughput experimental technologies are now routinely used to produce protein expression and interaction networks. When combined with complementary literature data, these networks become cellular pathways that are key elements for understanding the cell functions in context. Proteomic informatics enables the massive production of these data by storing them in dedicated databases allowing quality control, and by proposing adapted mining and visualization tools. Bioinformatics algorithms further allow prediction of protein networks by comparing genome sequences or by inferring networks across organisms. Even if one still lacks independent reference datasets and validation methods

to precisely evaluate the efficiency of such algorithms; they will probably soon be the main tools for looking for associations between heterogeneous biological data.

While the sheer amount of data made available by various means of analysis is a challenge in itself, its heterogeneity should be one of today's main concerns: not only does it make any hypothesis difficult to test extensively, it is only by cross-referencing independent data-sources that we will be able to develop a consistent corpus of knowledge and extract from it an adequate validation set for reliable comparison and accurate evaluation of existent and as yet undiscovered analysis methods.

ACKNOWLEDGEMENTS

We thank P. Durand, V. Schächter, Y. Chemama and P. Legrain for stimulating discussions and meticulous reading of the manuscript.

REFERENCES

Albert R, Jeong H, Barabasi AL. (2000). *Nature* **406**: 378–382.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, *et al*. (1997). *Nucleic Acids Res* **25**: 3389–3402.

Anderson L, Seilhamer J. (1997). *Electrophoresis* **18**: 533–537.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, *et al*. (2001). *Nucleic Acids Res* **29**: 37–40.

Bader GD, Hogue CW. (2000). *Bioinformatics* **16**: 465–477.

Bairoch A. (2000). *Nucleic Acids Res* **28**: 304–305.

Bartel PL, Roecklein JA, SenGupta D, Fields S. (1996). *Nature Genet* **12**: 72–77.

Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, *et al*. (2000). *Nucleic Acids Res* **28**: 73–76.

Dandekar T, Snel B, Huynen M, Bork P. (1998). *Trends Biochem Sci* **23**: 324–328.

Eilbeck K, Brass A, Paton N, Hodgman C. (1999). In *Seventh International Conference of Intelligent Systems for Molecular Biology*, pp. 87–94.

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. (2000). *Nature* **405**: 823–826.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. (1999). *Nature* **402**: 86–90.

Fields S, Song O. (1989). *Nature* **340**: 245–246.

Finley RL Jr, Brent R. (1994). *Proc Natl Acad Sci USA* **91**: 12980–12984.

Flajolet M, Rotondo G, Daviet L, Bergametti F, Inchauspé G, Tiollais P, *et al*. (2000). *Gene* **242**, 369–379.

Fromont-Racine M, Mayes AE, Brunet-Simon A, Rain JC, Colley A, Dix I, *et al*. (2000). *Yeast* **17**: 95–110.

Fromont-Racine M, Rain JC, Legrain P. (1997). *Nature Genet* **16**: 277–282.

Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, *et al*. (2002). *Nature* **415**: 141–147.

Gygi SP, Rochon Y, Franza BR, Aebersold R. (1999). *Mol Cell Biol* **19**: 1720–1730.

Haab BB, Dunham MJ, Brown PO. (2001). *Genome Biol* **2**:RESEARCH0004.

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. (2000). *Hum Mutat* **15**: 57–61.

Heinemeyer T, Chen X, Karas H, Kel AE, Kel OV, Liebich I, *et al*. (1999). *Nucleic Acids Res* **27**: 318–322.

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, *et al*. (2002). *Nature* **415**: 180–183.

Houry WA, Frishman D, Eckerskorn C, Lottspeich F, Hartl FU. (1999). *Nature* **402**: 147–54.
Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. (2001). *Proc Natl Acad Sci USA* **98**: 4569–4574.
Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, *et al*. (2000). *Proc Natl Acad Sci USA* **97**: 1143–1147.
Jenssen TK, Laegreid A, Komorowski J, Hovig E. (2001). *Nature Genet* **28**: 21–28.
Jeong H, Mason SP, Barabasi A-L, Oltvai ZN. (2001). *Nature* **411**: 41–42.
Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L. (2000). *Nature* **407**: 651–654.
Kanehisa M, Goto S. (2000). *Nucleic Acids Res* **28**: 27–30.
Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. (2000). *Nucleic Acids Res* **28**: 56–59.
Lee KH. (2001). *Trends Biotechnol* **19**: 217–222.
Legrain P, Jestin J-L, Schächter V. (2000). *Curr Opin Biotechnol* **11**: 402–407.
Legrain P, Wojcik J, Gauthier JM. (2001). *Trends Genet* **17**: 346–352.
Lemkin PF, Thornwall G. (1999). *Mol Biotechnol* **12**: 159–172.
Lueking A, Horn M, Eickhoff H, Bussow K, Lehrach H, Walter G. (1999). *Anal Biochem* **270**: 103–111.
MacBeath G, Schreiber SL. (2000). *Science* **289**: 1760–1763.
Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. (1999a). *Science* **285**: 751–753.
Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. (1999b). *Nature* **402**: 83–86.
Marcotte EM, Xenarios I, Eisenberg D. (2001). *Bioinformatics* **17**: 359–363.
Mayer ML, Hieter P. (2000). *Nature Biotechnol* **18**: 1242–1243.
McCraith S, Holtzman T, Moss B, Fields S. (2000). *Proc Natl Acad Sci USA* **97**: 4879–4884.
Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, *et al*. (2000). *Nucleic Acids Res* **28**: 37–40.
Mrowka R. (2001). *Bioinformatics* **17**: 669–671.
Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995). *J Mol Biol* **247**: 536–540.
Newman JR, Wolf E, Kim PS. (2000). *Proc Natl Acad Sci USA* **97**: 13203–13208.
Ono T, Hishigaki H, Tanigami A, Takagi T. (2001). *Bioinformatics* **17**: 155–161.
Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. (1999). *Proc Natl Acad Sci USA* **96**: 2896–2901.
Park J, Lappe M, Teichmann SA. (2001). *J Mol Biol* **307**: 929–938.
Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. (1997). *J Mol Biol* **271**: 511–523.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. (1999). *Proc Natl Acad Sci USA* **96**: 4285–4288.
Phizicky EM, Fields S. (1995). *Microbiol Rev* **59**: 94–123.
Pilpel Y, Sudarsanam P, Church GM. (2001). *Nature Genet* **29**: 153–159.
Quadroni M, James P. (1999). *Electrophoresis* **20**: 664–677.
Rain JC, Selig L, de Reuse H, Battaglia V, Reverdy C, Simon S, *et al*. (2001). *Nature* **409**: 211–216.
Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. (1999). *Nature Biotechnol* **17**: 1030–1032.
Sanchez C, Lachaize C, Janody F, Bellon B, Roder L, Euzenat J, *et al*. (1999). *Nucleic Acids Res* **27**: 89–94.
Schwikowski B, Uetz P, Fields S. (2000). *Nature Biotechnol* **18**: 1257–1261.

Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E. (1998). *Nucleic Acids Res* **26**: 43–45.
Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, *et al*. (1998). *Mol Biol Cell* **9**: 3273–3297.
Sprinzak E, Margalit H. (2001). *J Mol Biol* **311**: 681–692.
Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, *et al*. (2001). *Science* **13**: 13.
Tsoka S, Ouzounis CA. (2000). *Nature Genet* **26**: 141–142.
Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, *et al*. (2000). *Nature* **403**: 623–627.
Vidal M, Legrain P. (1999). *Nucleic Acids Res* **27**: 919–929.
Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, *et al*. (2000). *Science* **287**: 116–122.
Walter G, Konthur Z, Lehrach H. (2001). *Comb Chem High Throughput Screen* **4**: 193–205.
Whitelegge JP, le Coutre J, Lee JC, Engel CK, Prive GG, Faull KF, *et al*. (1999). *Proc Natl Acad Sci USA* **96**: 10695–10698.
Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, *et al*. (1999). *J Mol Biol* **289**: 645–67.
Wojcik J, Schächter V. (2001). *Bioinformatics* **17**: S296–S305.
Wu Q, Maniatis T. (1999). *Cell* **97**: 779–790.
Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. (2000). *Nucleic Acids Res* **28**: 289–291.
Yates JR III. (1998). *J Mass Spect* **33**: 1–19.

Bioinformatics for Geneticists. Edited by Michael R. Barnes and Ian C. Gray

ISBNs: 0-470-84393-4 (HB); 0-470-84394-2 (PB)

CHAPTER 17

Concluding Remarks: Final Thoughts and Future Trends

MICHAEL R. BARNES[1] and IAN C. GRAY[2]

[1]*Genetic Bioinformatics and* [2]*Discovery Genetics*
Genetics Research Division
GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK

The sequencing of the human genome is complete. This is an obvious milestone for all fields of biology, none more so than genetics. As we have seen throughout this book, the availability of a complete genome makes the study of the genetics of an organism much less haphazard, and bioinformatics is an essential enabling skill for the geneticist to make the most of the genome. In the pre-genome era, geneticists probed the genome like early explorers penetrating a dark continent ripe for exploration. Relying on only the most basic data they painstakingly reconstructed genes and methodically drafted maps to find disease alleles. Now, in the post-genome era, instead of stars and a compass the genetic explorers have the equivalent of a global positioning satellite system and a detailed A–Z directory of genes. With all this technology it might be hard to imagine how genetics can now fail to locate disease genes, but failure will still be a frequent outcome.

Why? Firstly, we may be looking for something that does not exist or is too small to detect using existing methodology. Complex human disease is a product of both environment and genes, but the environment is often overlooked as a source of disease, particularly in the current era of high-profile genetics. The contribution of a single gene to a multifactorial disease or trait may be vanishingly small and consequently even large studies may have insufficient power to detect it. Secondly our directory of genes may not be as comprehensive as we think, with significant weaknesses in certain areas, for example the assignment of function to poorly understood regulatory motifs and the degree and nature of inter-individual genome diversity. Thirdly our maps are not yet completely

error free. Bioinformatics cannot help with the first problem directly, although novel statistical methods may improve the chances of identifying small genetic effects and will form part of a continually evolving software suite for genetic analysis of complex traits. As more pieces of the genetic and environmental jigsaw puzzle are put into place for each complex trait, it should become progressively easier to position the remaining pieces to give a more complete picture. Although bioinformatics may be perceived as playing a secondary role in developing techniques for improved statistical analysis of complex trait data, it is the key to providing the equally important solutions required for a truly complete characterization of the genome coupled with unimpeachable data integrity.

17.1 HOW MANY GENES?

The biggest revelation of the human genome sequencing project was that humans appear to have fewer genes than we had expected. Estimates of the total number of human genes were widely anticipated to reach the 100,000 gene mark (Aparicio, 2000). As sequencing progressed these estimates were downgraded to 60–70,000 and finally as the first draft appeared estimates were consolidated to a mere 35,000 genes (Ewing and Green, 2000). If this figure is to be believed, then humans have only seven times as many genes as yeast, ▫ 2.5 times as many as the fly *Drosophila melanogaster* and less than twice as many as the nematode worm *Caenorhabditis elegans*. This figure may increase as understanding of the genome and gene prediction increases, although it seems unlikely that the number will rise beyond 50,000.

This smaller than expected number of genes might be viewed as good news for geneticists — fewer genes to screen for disease association. But fewer genes does not necessarily equate to reduced complexity. Complexity can manifest at many levels, including splicing, gene regulation, post-transcriptional editing and post-translational modification. In Chapter 12, we described the *Drosophila* DSCAM gene which has 115 exons which are alternatively spliced to code for 38,016 related but distinct protein isoforms (Schmucker *et al*., 2000). This remarkable gene gives us a hint that many of the gene models described so far in humans could under-represent the true diversity of the human gene repertoire. Instead it may be wise to view every gene transcript as a unit specific to a particular tissue, time or cellular condition. Alterations in any of these conditions could direct the expression of an alternative transcript.

It may also be pertinent to question the definition of a gene. Traditionally a gene is viewed as a protein-coding unit. Transcripts which do not obviously code for a protein are often dismissed as 'regulatory RNA' — a virtual dumping ground for transcripts which we are just beginning to understand (see Szymanski and Barciszewski, 2002). This situation is exacerbated by the wealth of data generated by genomics; for example a very large number of ESTs and cDNAs show no *in silico* evidence of splicing (i.e. by each end aligning either side of an intron in a genomic sequence). There are a number of explanations for the existence of such transcripts. They could be derived from a real gene but simply do not span an intron and therefore show no evidence of splicing; alternatively they could be *in vitro* artefacts generated during the construction of cDNA libraries or *in vivo* artefacts generated from cryptic promoters or pseudogenes.

This highlights one of the biggest challenges for the bioinformatic interpretation of the human genome — data overload. Gene prediction and annotation tools generally disregard unspliced ESTs as supporting evidence for the existence of a gene. This is a necessary precaution to avoid over-prediction of genes across the genome; tools designed to analyse whole genomes have to sacrifice sensitivity to avoid extensive over-prediction of genes

and to maintain the performance of genome analysis pipelines, but where geneticists seek to identify all candidate genes in a defined locus, it may be prudent to evaluate equivocal information such as unspliced ESTs in a more thorough fashion. This can be achieved easily with genome browser tools such as Ensembl and the UCSC human genome browser which present all available data across a locus. However, it is wise to proceed with caution when planning experimental work based on ambiguous data derived from *in silico* sources in order to avoid frustration as well as wasted time and resources. Simple, rapidly executed experiments to provide supporting evidence for the *in silico* observation should be the first step.

17.2 MAPPING THE GENOME AND GAINING A VIEW OF THE FULL DEPTH OF HUMAN VARIATION

Our incomplete understanding of genes and genome organization may not necessarily be a big problem for genetics. Experimental frameworks can be primarily focused on the physical and genetic composition of a region, in terms of genetic markers, recombination frequency and other characteristics, rather than its perceived functional content. 'Phenotype-driven' family-based whole genome linkage scans to identify genes responsible for monogenic traits illustrate one such approach. Use of linkage disequilibrium (LD) to identify genomic regions of genetic association is a second example, and is more appropriate for complex traits. This approach assumes little about the function of a marker or gene, but can allow mapping of a genetic association to a very small region (typically 10–100 kb) following the construction of detailed population-based LD maps. Completion of an LD map of the entire human genome will in itself be a highly significant milestone for genetics. Already provisional LD maps of chromosomes 21, 22 and 19 have been published (Dawson *et al*., 2002; Patil *et al*., 2001; Michael Phillips personal communication). A whole genome LD map generated by many of the former members of TSC should be made publicly available in late 2003. This will finally make comprehensive SNP-based whole genome association scans a realistic possibility; selecting SNPs which tag all of the major haplotype blocks across the genome will shift the emphasis toward good experimental design and away from conjecture when initiating genetic association studies.

However, evolution toward a whole-genome haplotype-based approach to genetic studies will present considerable challenges. For example, although all of the available evidence suggests that the majority of haplotypes in any given genomic region are common to multiple ethnic groups (Gabriel *et al*., 2002), haplotype frequencies may vary considerably between groups. Thus markers that tag common haplotypes in one ethnic group may not identify the most common haplotypes in other groups. Furthermore, approaches based on attempts to associate common haplotypes with a disease state are broadly reliant on the veracity of the 'common disease caused by common variants' hypothesis (see Pritchard, 2001). A low frequency haplotype which is associated with disease may evade detection, and a rare predisposing SNP occurring on a common haplotypic background may not be detected due to insufficient statistical power. Only empirical data gathered over the next few years will reveal the true scale of such issues. A further consideration is the increase in throughput and reduction in cost required to render the necessary scale of genotyping for population-based association studies, which are likely to require several million data points per genome-wide experiment, feasible. However significant investment in this area has led to promising improvements across a range of genotyping platforms over the last few years and we expect this trend to continue.

17.3 HOLISTIC ANALYSIS OF COMPLEX TRAITS

One of the weaknesses of genetic association studies is the difficulty in drawing a firm conclusion regarding the robustness of the finding from the statistical evidence for association between a given gene and trait, particularly if the level of significance is marginal. A key future application of bioinformatics is likely to be the drawing together of diverse threads of data from a number of sources in a more holistic approach toward the analysis of complex traits. The output from human linkage and population-based association studies can be combined with animal model quantitative trait loci, phenotypic data from systematic gene knock-out and transgenic mouse approaches, genome-wide expression data from microarrays, proteomic profiles and other sources, to provide a substantial body of evidence relating to the gene or locus in question. This will require the development of both new interfaces for the integration of disparate datasets and sophisticated global analysis software.

17.4 A FINAL WORD ON BIOINFORMATICS

It is always difficult to present a rapidly moving field such as bioinformatics in a book. Despite the best efforts of the authors, editors and publisher, by the time this book reaches the reader many of the tools described in the preceding chapters will have evolved to offer yet more functionality and utility. Keeping abreast of new developments in bioinformatics is as important an activity as using the data themselves. Current awareness of the field is essential to ensure that all of the relevant available data are captured, maximizing research efficiency. Finally, the best approach to becoming proficient in the use of software tools is often trial and error, and bioinformatics is no exception; trial and error *in silico* can obviate the far less desirable prospect of trial and error in the laboratory, so do not be afraid to experiment with bioinformatics applications — see what the human genome can yield in your hands. Good luck!

ACKNOWLEDGEMENTS

MRB and ICG would like to acknowledge the efforts of all of the authors who have contributed to this volume. This book has taken shape after many discussions with many of our collegues at GSK and in the wider scientific community. The first drafts were moulded into final chapters with the assistance of several willing proof readers, particularly Christopher Southan, Aruna Bansal, Ralph McGinnis and Mary Plumpton. We would also like to express our gratitude to Joan Marsh, Layla Paggett, Amie Tibble and Monica Twine at John Wiley for able assistance in the preparation of the manuscript. Finally this volume would not have been possible without the support and encouragement of Robin Dement and Ian Purvis at GSK.

REFERENCES

Aparicio SA. (2000). How to count human genes. *Nature Genet* **25**: 129–130.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, *et al*. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544–548.

Ewing B, Green P. (2000). Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet* **25**: 232–234.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, *et al*. (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225–2229.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al*. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.

Pritchard JK. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**: 124–137.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, *et al*. (2000). Drosophila DSCAM is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671–684.

Szymanski M, Barciszewski J. (2002). Beyond the proteome: non-coding regulatory RNAs. *Genome Biol* **3** (reviews 5): 1–8.