Bioinformatics and cancer target discovery

Brian Desany and Zemin Zhang

The convergence of genomic technologies and the development of drugs designed against specific molecular targets provides many opportunities for using bioinformatics to bridge the gap between biological knowledge and clinical therapy. Identifying genes that have properties similar to known targets is conceptually straightforward. Additionally, genes can be linked to cancer via recurrent genomic or genetic abnormalities. Finally, by integrating large and disparate datasets, gene-level distinctions can be made between the different biological states that the data represents. These bioinformatics approaches and their associated methodologies, which can be applied across a range of technologies, facilitate the rapid identification of new target leads for further experimental validation.

Brian Desany
Zemin Zhang*
Department of Bioinformatics
Genentech
1 DNA Way, M.S. 93
South San Francisco
CA 94080, USA
*email: zemin@gene.com

▼ Cancer is a disease characterized by extensive genomic abnormalities and aberrations in gene expression. Tumor-specific mutations, DNA amplifications and translocations can all distort the normal program of gene expression and function, resulting in misregulated activity of apoptosis, angiogenesis and cell proliferation. A class of new cancer therapies, including Avastin, Erbitux, Gleevec and Herceptin, exploit our knowledge of cancer biology by selectively inhibiting or killing cancer cells, while leaving normal cells unharmed. Most of these novel therapies specifically target cancer-causing gene products, such as Bcr–Abl, or proteins required for tumor progression and metastasis, such as vascular endothelial growth factor (VEGF). Other cancer targets, such as CD20 for Rituxan, are biomarkers and their physical presence confers vulnerability to cancer cells, although it is not clear whether they are causally implicated in oncogenesis.

Identification of additional cancer targets requires understanding the genetic events associated with cancer genesis and progression, as well as the identification of genes or gene mutations that either promote oncogenesis or are specifically associated with cancer. With the advent of genomic technologies and their application to the study of cancer biology, we are presented with exciting opportunities to perform HTS for novel cancer targets (Figure 1). Here we discuss the uses of bioinformatics methods and data sources in finding potential new cancer drug targets.

Sequence analysis as a broad-based selection method

Although a wide variety of protein classes are involved in cancer, targets of interest fall into three general categories: secreted proteins, cell surface receptors and markers, and intracellular kinases. Secreted and cell-surface proteins, such as VEGF and epidermal growth factor receptor (EGFR), are essential for intercellular communication, in particular signaling of cell proliferation and angiogenesis, and they are physically accessible to monoclonal antibodies, which have proven to be effective in treating cancers [1]. By contrast, kinases are pivotal players in signal transduction and provide assayable enzymatic activities for screening small-molecule inhibitors. A recent survey of human cancer genes indicated an over-representation of protein kinases in a collection of genes that are causally implicated in cancer [2].

Homologues of proteins involved in cancer-related pathways might have related biological roles. With the human genome completely sequenced, and the availability of pair-wise sequence alignment programs such as BLAST [3] or hidden Markov model-based programs such as the HMMER package [4], finding sequence homologues of known cancer genes has become relatively straightforward. However, new cancer targets might not bear sequence similarity to known cancer genes. It is therefore necessary to cast a wider net to include molecules with more general structural features

1359-6446/04/$ – see front matter PII: S1359-6446(04)03224-6

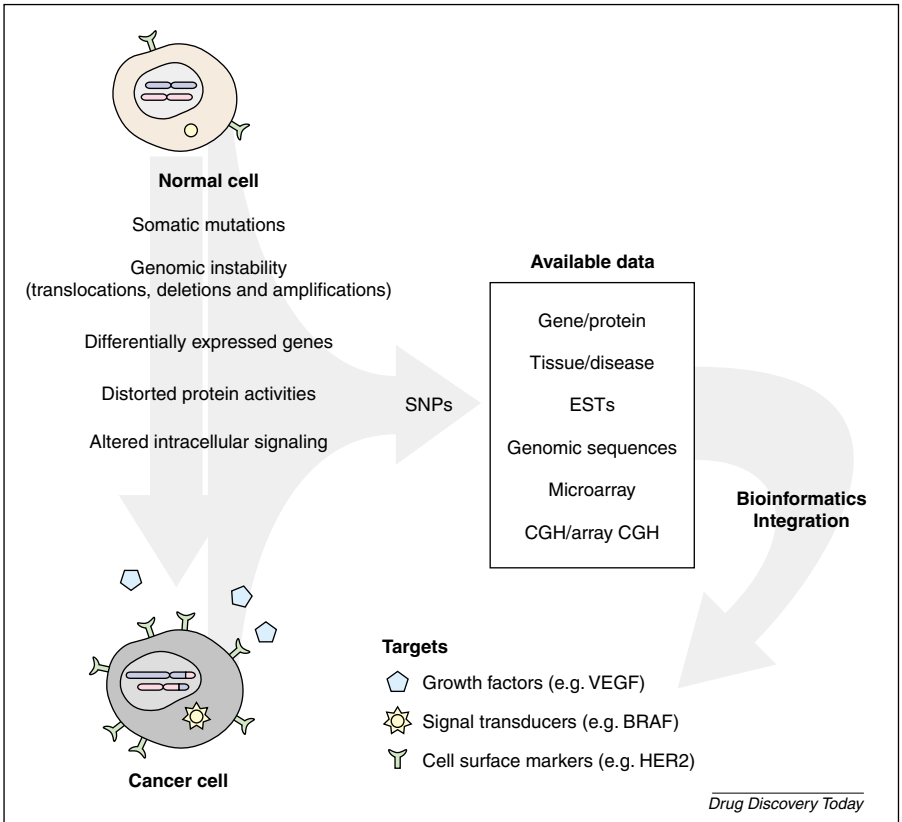

Figure 1. Bioinformatics integration of heterogeneous biological data provides opportunities for understanding cancer genetics and identifying therapeutic targets. Cancer cells are genetically different from normal cells due to somatic point mutations and aberrations at the genomic level (translocations, deletions and amplifications). The phenotypic consequences of these genetic changes include overexpression of cell-surface markers and activated kinases, presenting cancer-specific targets for therapeutic treatment. Various types of data from expressed sequence tag (EST), microarray and comparative genomic hybridization (CGH) technologies can be combined to detect point mutations, DNA copy number changes and differentially expressed genes in cancer cells, leading to the identification of cancer-causing genes as well as other genetic properties associated with cancer.

of existing cancer targets, such as the presence of a signal peptide, a transmembrane domain or a protein kinase domain [5]. This can be achieved using a variety of computational prediction tools, including SignalP for signal peptides [6] and TMHMM for transmembrane domains [7]. The number of genes found in this way is expected to be large, and a majority of these genes might not be directly associated with cancer. One way to trim down the large pool of initial selection is to use biological assays to identify cancer-associated genes. This experimentally driven process is illustrated by the successful identification of EG-VEGF, which was initially found by sequence analysis to be a novel secreted protein with no predictable function and was subsequently shown to possess strong mitogenic activity in a cell-based assay [8]. A more general, efficient and powerful approach is to use the bioinformatics methods described below to maximize our functional knowledge by exploiting high-throughput genomic data sources, such as gene expression assays and mutational studies, reducing the need for a large prior experimental investment.

Gene expression profiling using microarrays

Many techniques have been developed for using microarray gene expression data to study various facets of cancer biology. Here we address some of the more common computational approaches that have found applications in target discovery and refer the reader to a recent review [9] for a general overview of microarray technology, experimental design and data processing issues.

Microarray analytical approaches generally fall into two categories: supervised and unsupervised learning. Supervised learning uses the known identities of samples and the expression patterns of a limited number of genes to generate predictors that are used to classify unknown samples for diagnostic purposes [10,11]. Unsupervised learning involves using a distance metric to group samples or genes together into related groups, or clusters [12], and has direct applications to target discovery. Candidates for further validation as potential targets can be defined by gene clusters that have desirable expression characteristics, for example, more highly expressed in tumors than in normal cells.

The most straightforward inference of differential expression can be made from a fold-change measurement, which is the ratio of the expression levels of a given gene between two samples. Interesting gene clusters would therefore contain genes with a high fold-change between tumor and normal cells. Although useful and intuitive, fold-change suffers by not taking measurement noise or biological variability into account, particularly because cancer is inherently a heterogeneous disease. It has become common to include as many tumor samples as possible in a given analysis so that statistical methods, such as ANOVA and t-like tests [13], can be used on a gene-by-gene basis to compare groups of large numbers of samples to each other, rank gene clusters and assign statistical significance to the results.

Biological variability is partially mitigated by increasingly precise sample acquisition techniques, such as laser capture microdissection, that can reduce the contribution of non-relevant cell types to the expression profiles. However, cancers with similar histopathological features, and even the same diagnosis, can actually represent more than a single disease entity at the molecular level. Staudt and colleagues have used hierarchical clustering to subclassify diffuse large B-cell lymphoma into three subtypes, each characterized by specific groups of gene signatures [14]. In addition to shedding light on the biology of diffuse large B-cell lymphoma (DLBCL), this underscores the fact that even using techniques with high statistical power, the quality of unsupervised clustering and differential expression analysis will depend largely on the underlying assumptions about sample identity.

Genes with similar expression patterns are likely to have related functions ('guilt by association' [15]), therefore, finding genes with expression patterns similar to those of known tumor markers is usually a convenient and effective method to employ in the search for new candidates. This can be done by comparing the gene expression patterns in a dataset to the pattern of the known marker in a pair-wise fashion using an appropriate distance metric (e.g. Pearson's correlation coefficient) and choosing the highest-ranking hits. Even considering more comprehensive unsupervised clustering approaches, this focused alternative way of 'grabbing the low-hanging fruit' still has merit, especially when applied to data from new chips representing genes not present on prior chips.

Equally important to these various methodologies is the accumulation of microarray data suitable for cancer target discovery. Numerous gene expression studies in many cancer types are associated with entire datasets that can be downloaded, many in public repositories specifically dedicated to the dissemination of these valuable data. These include websites such as the Stanford Microarray Database (http://genome-www5.stanford.edu/MicroArray/SMD), the NCBI's Gene Eexpression Omnibus repository (http://www.ncbi.nlm.nih.gov/geo), the EBI's ArrayExpress (http://www.ebi.ac.uk/arrayexpress) and the MIT Cancer Genomics Program (http://www.broad.mit.edu/cancer/). These are valuable resources for target discovery. For example, the Global Cancer Map data collection [16], which contains data for 218 tumor samples of 14 common tumor types along with 90 normal tissues, can be readily utilized for finding genes highly expressed in lung cancer but low in other normal tissues. There are still tremendous challenges in integrating microarray data from different sources, as microarray platform and data representation can be heterogeneous and inconsistent, but the standardization [17] and integration [18] of microarray data should emerge in the coming years to enable uniform representation of array data and cross-platform data comparison for effective target discovery purposes.

Finally, attractive targets are not defined solely by their expression patterns. Other criteria, such as type of molecule (e.g. kinase), subcellular localization (e.g. cell surface) and biological pathway (e.g. angiogenesis), are important in the decision to follow-up on a potential new target. To this end, resources such as the Gene Ontology Project (http://www.geneontology.org) and the Kyoto Encyclopedia of Genes and Genomes Pathways Project (http://www genome.ad.jp/kegg) attempt to place genes in the context of biological function, location and pathway. The recent W eb Server issue of *Nucleic Acids Research* [19] contains a current overview of many tools that use these and other resources.

Digital expression profiling using EST and SAGE

Gene expression profiling is not necessarily synonymous with microarrays. 'Digital expression' based on either expressed sequence tags (ESTs) or serial analysis of gene expression (SAGE) is complementary to microarrays and can be just as powerful. Both EST-derived expression and SAGE are based on the principle that the frequency of sequence tags sampled from a pool of cDNAs is directly proportional to the expression level of the corresponding gene (see Figure 2). Digital expression offers three major advantages over microarrays. First, the simple digital data format in sequence clone counts and frequencies enables direct and platform-independent data comparison among different data collections from multiple tissues. Second, because there is no need for designing any DNA chips, no prior knowledge of gene sequence is required and therefore many novel genes not covered by microarrays are represented. Finally, because the expression level is represented by mRNA abundance relative to all transcripts and is independent of probe selection and hybridization biases, digital expression can be a more quantitative measurement of gene expression than microarrays [20].

Although ESTs were initially collected primarily for novel gene identification [21], their value in gene expression analysis became obvious with the accumulation of EST data and development of appropriate computational tools. There are currently over 5 million human EST sequences available in the public database [22], derived from many laboratories and tissue types over the last ten years. Notably, a significant fraction of these ESTs are derived from cancer tissues as a result of the large-scale efforts of the Cancer Genome Anatomy Project (CGAP; http://www.ncbi.nlm.nih.gov/ncicgap) at the National Cancer Institute

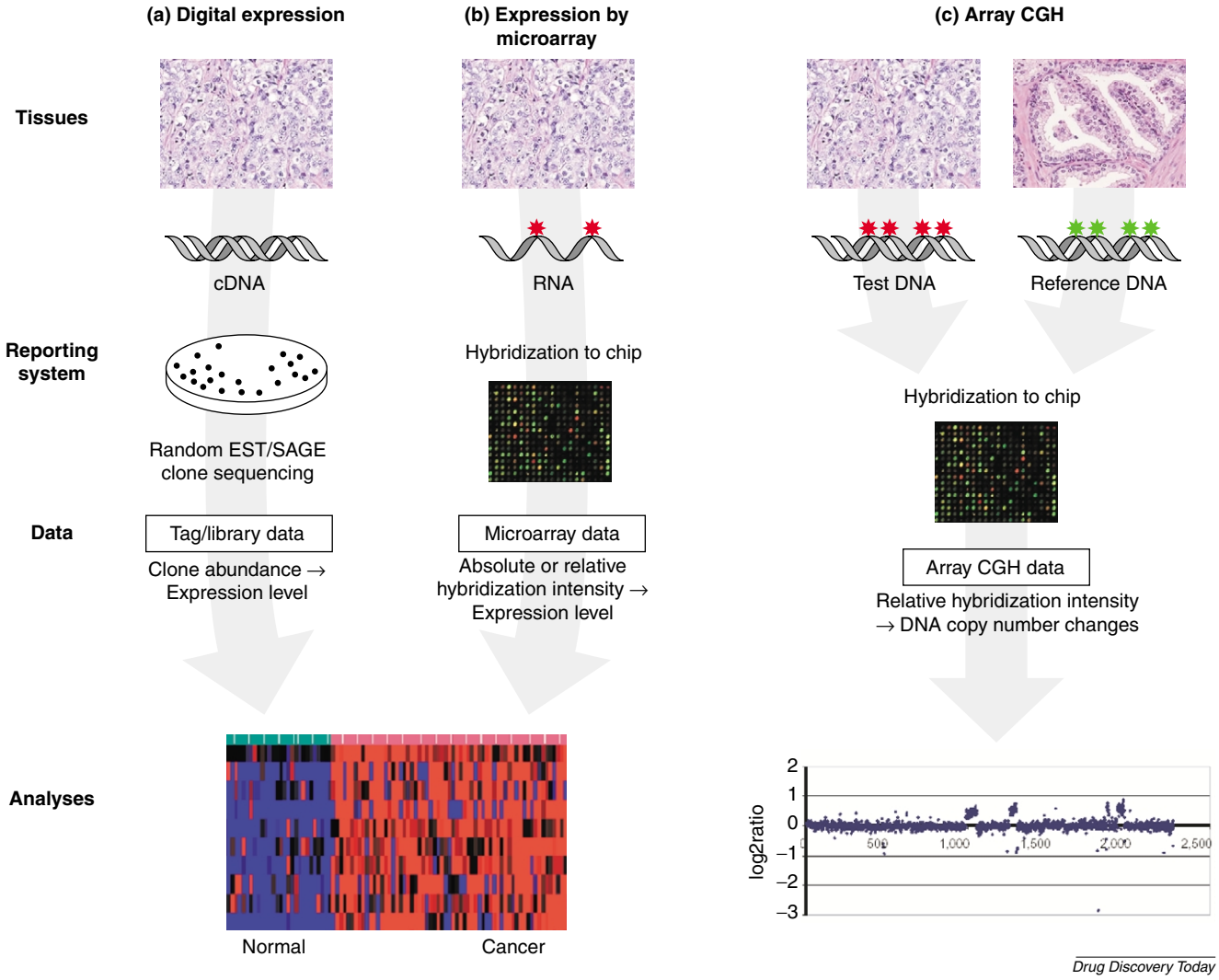

Figure 2. A schematic representation of several types of high-throughput data useful for cancer bioinformatics studies. **(a)** For digital expression, collections of random clones harboring EST or SAGE tags are picked and DNA sequences determined. The expression level of a particular gene is measured by the relative frequency of sequence tags for the gene. **(b)** Microarrays measure transcript abundance by hybridizing fluorescently labeled RNA to DNA probes on a chip. Gene expression can be measured either on an absolute intensity scale or as a ratio of intensities between a test (cancer) and a reference (normal) sample. Both digital and microarray expression data can be used for large-scale differential gene expression analysis. **(c)** Array comparative genomic hybridization (CGH) is similar to expression microarrays, except that instead of RNA, labeled genomic DNA is hybridized to DNA chips. In this case, the relative hybridization intensity of a cancer DNA sample to a normal DNA sample (the reference) represents the relative abundance of genomic DNA in the cancer sample, which is a measure of copy number change.

(NCI) to generate EST libraries from tumor samples [23]. Excluding normalized or subtracted cDNA libraries that are not suitable for expression studies, ESTs present an attractive source for not only differential expression analysis between normal and cancer tissues but also for multi-tissue expression profiling needed for toxicity prediction.

One common application of EST data is the pair-wise comparison of libraries (or pools of libraries) of different origins, for example, tumor versus normal libraries. Such comparisons are facilitated by tools like Digital Differential Display (DDD) [24] at the NCBI, cDNA Digital Gene Expression Displayer (DGEN), and xProfiler at CGAP. The reliability of this kind of expression result is largely dependent on the size of the component libraries and can be measured by Fisher's exact test [24], the Z-test for two sample proportions, and other statistical measurements [25].

Although pair-wise expression comparison is useful in finding genes that are expressed at a higher level in tumor than the corresponding normal tissues, cancer target discovery requires comprehensive profiling of gene expression in

Table 1. Representative online resources for cancer target finding and analysis

Name	Description	URL
Cancer Genes	Collection of cancer genes based on mutation data	http://www.sanger.ac.uk/genetics/CGP/Census/
Stanford Microarray Database (SMD)	Repository of microarray data with analysis tools	http://genome-www5.stanford.edu/
Whitehead Institute Center for Genome Research, Cancer Genomics Program	Repository of microarray data from cancer genomics publications	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Gene Expression Omnibus (GEO)	Repository of microarray data from a variety of sources, including CGH	http://www.ncbi.nlm.nih.gov/geo/
Gene Ontology Project (GO)	Controlled vocabulary describing molecular function, biological process, and cellular component	http://www.geneontology.org/
Kyoto Encyclopedia of Genes and Genomes Pathways (KEGG Pathways)	Molecular interaction networks, including metabolic and regulatory pathways, and molecular complexes	http://www.genome.ad.jp/kegg/metabolism.html
SAGEmap	SAGE data repository and analysis tools	http://www.ncbi.nlm.nih.gov/sage
SAGE Genie	Analytical method of tag-to-gene mapping with intuitive display	http://cgap.nci.nih.gov/SAGE
DGEN	Digital gene expression display tool that quantitatively compares two pools of EST libraries	http://cgap.nci.nih.gov/Tissues/GXS
cDNA xProfiler	Compares gene expression of two pools of EST libraries	http://cgap.nci.nih.gov/Tissues/xProfiler
GEPIS	EST-based gene expression profiling data and analysis	http://www.cgl.ucsf.edu/Research/gene.ntech/gepis/
Progenetix	Repository of cytogenetic abnormalities in human cancer	http://www.progenetix.com/
SKY/M-FISH and CGH Database	Repository of cytogenetic abnormalities in human cancer	http://www.ncbi.nlm.nih.gov/sky/
SNP500Cancer	Validated SNPs in cancer genes	http://snp500cancer.nci.nih.gov/home.cfm

a wide spectrum of normal and tumor tissues. To facilitate this process, we have recently built a computational tool named GEPIS (gene expression profiling *in silico*), which profiles expression levels in multiple normal and tumor tissues based on pools of EST libraries [20]. By iterative examination of such profiles for all available genes, we used GEPIS to find desirable expression patterns – high expression in one type of tumor tissue but low elsewhere, including the non-related normal tissue types. The expression patterns of many of the targets identified by GEPIS have been subsequently validated by other experimental methods, including quantitative RT–PCR and tissue microarrays. Alternatively, as with microarray data, one could use guilt-by-association methods to identify additional targets whose expression patterns mimic those of known cancer-associated genes [26].

Like EST-based expression analysis, SAGE relies on the frequency of occurrence of sequence tags in a library, but in this case the tags are 14-bp or 21-bp short sequences that identify the cDNAs. The SAGE method enzymatically concatenates multiple such short sequence tags into each clone, enabling many data points to be obtained from a single sequencing read [27]. Subsequent analysis is similar to EST-based methods. CGAP has sponsored the generation of SAGE libraries in numerous normal and cancerous tissues and makes the data available at the NCBI and CGAP websites. Coupled with computational tools, such as SAGEmap [28] from NCBI and SAGE Genie [29] from CGAP, SAGE libraries have become an excellent source for cancer-related expression analysis. SAGE data mining followed by PCR screening was able to identify candidate glioblastoma markers and antigens [30], and there are

many other examples of the application of SAGE in cancer target discovery [31]. Given the common principle of EST- and SAGE-based expression and their digital data format, computational tools can be developed that combine sequence tags between platforms to produce comprehensive profiles that depict expression levels in thousands of diverse tissues. The continual accumulation of ESTs and SAGE tags, which currently stands at over 10 million, will only rise in value as a resource for cancer target discovery.

Cancer association from recurrent DNA amplification

Part of the multi-step process of tumor formation is a period of genomic instability usually resulting in regions of genomic copy number increase. Oncogenes such as c-myc have long been known to be associated with regions of high copy number, and mapping such amplicons in tumor cells has become a common method for searching for new oncogenes or determining which known oncogenes might be contributing to a particular cancer type. The methods for detecting recurrent DNA amplifications have traditionally been cytogenetic [comparative genomic hybridization (CGH), M-FISH, SKY], and have resulted in the characterization of many cancer-associated amplicons, albeit with resolution limited to what is detectable in a microscope (e.g. the order of 2–10 Mb). Databases such as Progenetix.net [32] and the NCI and NCBI's SKY/M-FISH and CGH Database have begun to serve as public repositories for regions of chromosomal amplifications and deletions.

Recently, microarray technology has been adapted for CGH (array-CGH, or matrix-CGH) to increase the resolution over the microscope-based approaches and to enable much finer mapping of cancer amplicons [33]. Array-CGH is similar in principle to array gene expression (see Figure 2). The probes on the chip are mapped sequences, usually BAC clones, although cDNAs and oligonucleotides have also been used. The hybridizing test samples are genomic DNA from tumor samples. Genomic DNA from a normal sample serves as a universal reference, with the test:reference ratio representing the relative copy number of a given probe. With the emergence of array-CGH as a high-throughput technology, and particularly as the chips become denser and resolution improves [34], bioinformatics approaches have enabled the identification of amplicons that occur in many tumors. The area of minimal overlap of such recurrent amplicons greatly facilitates identification of the oncogene candidates in those tumors [35].

Like many genomic technologies, array-CGH is particularly powerful when combined or integrated with complementary data from independent genomic approaches. Gene expression tends to be higher for genes located within amplicons [36,37], and array-CGH is often used in conjunction with expression microarrays to identify likely 'driver' genes within amplicons [38]. Chip design is an important consideration for these studies because there is also a subtle correlation between expression levels and probe localization on chips [39]. In another approach, Zardo *et al.* [40] have searched for potential tumor suppressor genes in brain tumors by using restriction landmark genomic scanning (RLGS) to identify areas of high methylation, which were then compared to regions of genomic copy number loss obtained through array-CGH analysis. Protein expression on tissue microarrays [41], SKY [42] and conventional CGH have all been used to supplement array-CGH data and narrow down the field of potential oncogenes.

Gene expression data alone can also be used independently to discover regions of genome amplification *in silico*. In an EST-based whole-genome transcriptome analysis, the Z-test was used to measure the extent of up-regulation in cancer for a single gene, and a sensing index was designed to scan for genomic windows where a cluster of genes show higher expression in cancer [43]. The resulting non-random Regions of Increased Tumor Expression (RITEs) were found in all tissues examined, and they appear to correlate with experimentally determined amplicons. Similarly, when microarray-based expression data were analyzed in the context of genomic organization, malignancy-associated regions of transcriptional activation (MARTA) were detected as spatial clusters of highly expressed genes in various cancers [44]. In the absence of array CGH data, both RITE and MARTA data provide rapid and valuable cancer amplicon leads, which tend to harbor genes that are causally implicated in cancer.

Cancer gene finding from variant analysis

As a genetic disease, cancer arises due to the accumulation of mutations in crucial genes that influence cell proliferation, differentiation and death. Identification of mutated genes that are causally implicated in oncogenesis is an important aspect of target discovery. Either loss-of-function mutations in tumor suppressors like p53 [45,46] or gain-of-function mutations in oncogenes like *BRAF* [47] can play promoting roles in oncogenesis. Identification of mutations that occur predominately in cancer should help expand the collection of tumor suppressors and oncogenes, which are often targeted for cancer therapy. In addition, knowledge of cancer-causing mutations will facilitate early diagnosis and help identify groups of patients who respond better to a particular cancer treatment. Recently, studies of EGFR somatic point mutations in lung cancer patients underscore the link between clinical responses to mutations in target genes [48,49].

Two types of cancer-specific genetic changes are of particular interest: cancer-associated point mutations and splicing variants. Although many point mutations have no functional consequence, if one is found to be statistically associated with the cancer phenotype it might then be considered to be causal in cancer development. This is true even if the functional consequence of the variant/mutation is not immediately obvious, although non-silent mutations are clearly the most attractive candidates for follow-up. The same rationale applies to splice variants that are predominately and non-randomly observed in cancer samples. Compared with somatic point mutations, single nucleotide polymorphisms (SNPs) are more frequent, genetically stable, and often do not cause any deleterious effect. Some SNPs might cause higher susceptibility to cancer and are therefore expected to be overrepresented in cancer samples. Here, we do not distinguish somatic mutations from SNPs and refer to them collectively as cancer SNPs.

By linking tissue source, gene and sequence information together, ESTs provide opportunities for *in silico* detection of SNPs and splicing variants. Sequence alignments between ESTs and reference genes are often used for the initial detection of SNPs [50], and nearly 3 million SNPs have been deposited into the NCBI's dbSNP database [51]. With EST libraries assigned with either normal or tumor sources, statistical analyses such as Fisher's exact test can be used to determine whether a specific SNP is over-represented in cancer [52]. Again, the same principles apply to the study of cancer-specific splice variants. Although exon junction microarrays are starting to show great utility [53], most large-scale RNA splicing analyses rely on intron/exon structures revealed by EST sequences [54]. With EST source information, every available human mRNA can be compared with its corresponding normal and cancer ESTs using sequence alignment software, and tumor-associated splice forms were identified using the Z-test [55]. Alternatively, with normal and cancer ESTs aligned to genomic sequences, log odd ratio (LOD) can be calculated to detect splicing forms overrepresented in cancer tissues [56]. It is worth noting that although the *in silico* detection of gene variants holds great promise, it is subject to the same limitations of all bioinformatics approaches, which is that the results need experimental validation to avoid false leads derived from noisy data. Nevertheless, the high-throughput nature of *in silico* screening provides a valuable initial step in finding mutated genes that are causally involved in cancer.

Conclusion

As a genetic disease, cancer leaves a trail of genetic markers accompanying tumorigenesis and cancer progression. Somatic mutations, genomic instability and altered gene expression patterns all provide possible ways to distinguish cancer from normal cells, and such distinctions can help us develop therapies that specifically target cancer cells. As an enabling technology, bioinformatics has evolved in many ways that enable us not only to identify players in cancer pathways but also to comprehend genetic changes in cancer.

Cancer target discovery requires integrated and high-throughput analysis of genes, gene variants, expression and DNA copy number changes. A recurrent theme for bioinformatics data mining is that data from many different sources continues to accumulate in public repositories. As the amount and variety of the data continues to increase, bioinformatics methods can continue to be developed, refined and applied to exploit the larger datasets. Bioinformatics is not limited to a specific technology or type of information, and in fact much of its utility lies in integrating disparate data into a web of evidence used to weigh the quality of potential targets. The application of bioinformatics methods in cancer target discovery is starting to generate many exciting target leads for further experimental validation. Although bioinformatics-driven target discovery is still in its infancy, it has already become an indispensable piece of technology for cancer therapy development in this post-genomic era.

Acknowledgements

The authors would like to thank Colin Watanabe, Thomas Wu and Paul Polakis for critical review of the manuscript and Allison Bruce for assistance in preparing the figures.

References

1 Green, M.C. *et al.* (2000) Monoclonal antibody therapy for solid tumors. *Cancer Treat. Rev.* 26, 269–286
2 Futreal, P.A. *et al.* (2004) A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183
3 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
4 Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* 14, 755–763
5 Clark, H.F. *et al.* (2003) The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res.* 13, 2265–2270
6 Nielsen, H. and Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc. Int. Conf. Intel. Syst. Mol. Biol.* 6, 122–130
7 Krogh, A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580
8 LeCouter, J. *et al.* (2001) Identification of an angiogenic mitogen selective for endocrine gland endothelium. *Nature* 412, 877–884
9 Leung, Y.F. and Cavalieri, D. (2003) Fundamentals of cDNA microarray data analysis. *Trends Genet.* 19, 649–659
10 Golub, T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537

11 Wu, T.D. (2002) Large-scale analysis of gene expression profiles. *Brief. Bioinform.* 3, 7–17
12 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
13 Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121
14 Alizadeh, A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511
15 Quackenbush, J. (2003) Genomics. Microarrays – guilt by association. *Science* 302, 240–241
16 Ramaswamy, S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U. S. A.* 98, 15149–15154
17 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 29, 365–371
18 Detours, V. *et al.* (2003) Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett.* 546, 98–102
19 Editors (2004) Editorial. *Nucleic Acids Res.* 32, W1
20 Zhang, Y. *et al.* GEPIS - quantitative gene expression profiling in normal and cancer tissues. *Bioinformatics* (in press)
21 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656
22 Boguski, M.S. *et al.* (1993) dbEST–database for 'expressed sequence tags'. *Nat. Genet.* 4, 332–333
23 Brentani, H. *et al.* (2003) The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13418–13423
24 Scheurle, D. *et al.* (2000) Cancer gene discovery using digital differential display. *Cancer Res.* 60, 4037–4043
25 Stekel, D.J. *et al.* (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.* 10, 2055–2061
26 Walker, M.G. *et al.* (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.* 9, 1198–1203
27 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
28 Lash, A.E. *et al.* (2000) SAGEmap: a public gene expression resource. *Genome Res.* 10, 1051–1060
29 Boon, K. *et al.* (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11287–11292
30 Loging, W.T. *et al.* (2000) Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Res.* 10, 1393–1402
31 Porter, D. and Polyak, K. (2003) Cancer target discovery using SAGE. *Expert Opin. Ther. Targets* 7, 759–769
32 Baudis, M. and Cleary, M.L. (2001) Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics* 17, 1228–1229
33 Snijders, A.M. *et al.* (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* 29, 263–264
34 Ishkanian, A.S. *et al.* (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* 36, 299–303
35 Albertson, D.G. *et al.* (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.* 25, 144–146
36 Pollack, J.R. *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12963–12968
37 Virtaneva, K. *et al.* (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl. Acad. Sci. U. S. A.* 98, 1124–1129
38 Martinez-Climent, J.A. *et al.* (2003) Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations. *Blood* 101, 3109–3117
39 Kluger, Y. *et al.* (2003) Relationship between gene co-expression and probe localization on microarray slides. *BMC Genomics* 4, 49
40 Zardo, G. *et al.* (2002) Integrated genomic and epigenomic analyses pinpoint biallelic gene inactivation in tumors. *Nat. Genet.* 32, 453–458
41 Zaharieva, B.M. *et al.* (2003) High-throughput tissue microarray analysis of 11q13 gene amplification (CCND1, FGF3, FGF4, EMS1) in urinary bladder cancer. *J. Pathol.* 201, 603–608
42 Harding, M.A. *et al.* (2002) Functional genomic comparison of lineage-related human bladder cancer cell lines with differing tumorigenic and metastatic potentials by spectral karyotyping, comparative genomic hybridization, and a novel method of positional expression profiling. *Cancer Res.* 62, 6981–6989
43 Zhou, Y. *et al.* (2003) Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. *Cancer Res.* 63, 5781–5784
44 Glinsky, G.V. *et al.* (2003) Common malignancy-associated regions of transcriptional activation (MARTA) in human prostate, breast, ovarian, and colon cancers are targets for DNA amplification. *Cancer Lett.* 201, 67–77
45 Levine, A.J. *et al.* (1991) The p53 tumour suppressor gene. *Nature* 351, 453–456
46 Hollstein, M. *et al.* (1991) p53 mutations in human cancers. *Science* 253, 49–53
47 Davies, H. *et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature* 417, 949–954
48 Lynch, T.J. *et al.* (2004) Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med.* 350, 2129–2139
49 Paez, J.G. *et al.* (2004) EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500
50 Brett, D. *et al.* (2000) EST analysis online: WWW tools for detection of SNPs and alternative splice forms. *Trends Genet.* 16, 416–418
51 Sherry, S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311
52 Qiu, P. *et al.* (2004) Genome wide *in silico* SNP-tumor association analysis. *BMC Cancer* 4, 4
53 Johnson, J.M. *et al.* (2003) Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* 302, 2141–2144
54 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859
55 Wang, Z. *et al.* (2003) Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Res.* 63, 655–657
56 Xu, Q. and Lee, C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.* 31, 5635–5643