Online Tools for Bioinformatics Analyses in Nutrition Sciences\(^1,2\)

Sridhar A. Malkaram,\(^3,5\) Yousef I. Hassan,\(^4,5\) and Janos Zempleni\(^3\)*

\(^3\)Department of Nutrition and Health Sciences, University of Nebraska, Lincoln, Nebraska; \(^4\)Nutrition and Food Science Department, Faculty of Health Sciences, University of Kalamoon, Deirattiah, Syria

ABSTRACT

Recent advances in “omics” research have resulted in the creation of large datasets that were generated by consortia and centers, small datasets that were generated by individual investigators, and bioinformatics tools for mining these datasets. It is important for nutrition laboratories to take full advantage of the analysis tools to interrogate datasets for information relevant to genomics, epigenomics, transcriptomics, proteomics, and metabolomics. This review provides guidance regarding bioinformatics resources that are currently available in the public domain, with the intent to provide a starting point for investigators who want to take advantage of the opportunities provided by the bioinformatics field. Adv. Nutr. 3: 654–665, 2012.

Introduction

The Human Genome Project released an advanced version of the human genome in 2003 in 2 landmark papers (1,2). The availability of high-quality, comprehensive sequence data broke ground for a new era of research in various disciplines, including human nutrition and its role in disease prevention. Nutrigenomics emerged alongside other omics fields after the genome revolution and deals with the study of nutrient–gene interactions that could give way to possible dietary interventions in the overall goal of maintenance of optimum health or prevention of disease (3). The societal benefits of dietary interventions within the nutrigenomics framework are evident by looking at success stories in nutrition textbooks. For example, the implementation of newborn screening programs for early detection of phenylketonuria and for biotinidase deficiency, combined with the dietary restriction of phenylalanine and supplementation with biotin, respectively, has resulted in excellent prognoses for afflicted individuals.

Although dietary interventions in individuals with well-defined gene mutations have long been part of routine medical practice, they represent only the tip of the iceberg compared with the potential societal and individual benefits of taking full advantage of another recent milestone in genomics, i.e., the release of the “1000 Genomes” report (4). This report, which identifies ~15 million single nucleotide polymorphisms (SNP)\(^6\), copy number variations, deletions, and other interindividual variations in the human genomes from 3 projects, predicts a 5-fold greater frequency of sequence variations compared with previous estimates. One can assume with a reasonable level of confidence that many of these SNP are linked to disease risk. Dietary intervention offers an effective and cost-efficient approach to prevent disease (5). One of many classic examples in support of this theory relates to studies of SNP in the human methylene tetrahydrofolate reductase (\(\text{MTHFR}\)) gene; these SNP may predispose individuals to an increased risk of heart attack, renal disease, and birth defects (6,7).

Based on the definition of genetics and genomics by the WHO (8), nutrigenetics can be defined as studies of nutrition and heredity, whereas nutrigenomics is the study of the mutual interactions among dietary molecules, genes, and gene function. The main difference between genomics and genetics is that genetics scrutinizes the functioning

---

\(^1\) Supported by a contribution of the University of Nebraska Agricultural Research Division, supported in part by funds provided through the Hatch Act. Additional support was provided by National Institutes of Health grants DK063945, DK077816, and DK082476.

\(^2\) Author disclosures: S. A. Malkaram, Y. I. Hassan, and J. Zempleni, no conflicts of interest.

\(^3\) These authors contributed equally to the manuscript.

\(^*\) To whom correspondence should be addressed: E-mail: jzempleni2@unl.edu.

\(^6\) Abbreviations used: ChIP, chromatin immunoprecipitation; dbSNP, Single Nucleotide Polymorphism Database; EBI, European Bioinformatics Institute; EMBL, European Molecular Biology Laboratory; EST, expressed sequence tags; GEO, Gene Expression Omnibus; GPMDB, Global Proteome Machine database; GWAS, genome-wide association studies; HMDB, Human Metabolome Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; MTHFR, methylene tetrahydrofolate reductase; NCBI, National Center for Biotechnology Information; SNP, single nucleotide polymorphism; UCSC, University of California, Santa Cruz.
and composition of the single gene, whereas genomics addresses all genes and their interrelationships to identify their combined influence on the growth and development of the organism (8). In this review, nutrigenomics refers to both nutrigenetics and nutrigenomics.

Nutrition researchers, particularly those in the nutrigenomics area, need to make a whole-hearted and concerted effort to integrate bioinformatics expertise in their toolboxes to obtain a valuable partner in the genetics and genomics arena and to take full advantage of the many new opportunities that have emerged since the sequencing of the human genome. This review introduces the reader to publicly available datasets pertinent to nutrigenomics research and to some of the basic, yet valuable, online analysis tools for processing and interrogating such datasets. In particular, nutrigenomics core competencies in genomics, epigenomics, transcriptomics, proteomics, and metabolomics are highlighted (Table 1).

### Genomes

**Genomics and associated databases**

Access to accurate and complete genome sequences is a fundamental prerequisite for conducting genomics research. DNA sequencing has seen many breakthrough technological developments, from automated capillary-based methods to the recently developed ultrahigh-throughput methods. This field continues to develop rapidly and is now moving toward single-molecule sequencers (9). Sequence databases can be classified into 2 clusters, namely, primary sequence depositories and secondary databases derived from the primary databases (see the genome variations section) (Table 2). The primary sequence depositories store the raw sequence data obtained from various independent sequencing experiments and are generally considered starting points for subsequent research. Depositories such as GenBank (10), European Molecular Biology Laboratory (EMBL) Bank (11), and DNA DataBank of Japan (12) maintain identical data through a daily and mutual exchange of data. The sequences are grouped by taxonomic divisions (such as bacteria, fungi, invertebrates, and vertebrates) and into various data classes [expressed sequence tags (EST), genome survey sequences, sequence tagged sites, transcriptome shotgun assembly, environmental sequences, synthetic sequences, high-throughput cDNA sequences, high-throughput genomic sequences, and whole genome shotgun sequences] based on the type of sequencing experiment (13). In addition to storing the raw and annotated general nucleotide sequences, each of the 3 centers maintain a sequence read archive and a trace archive, which store raw sequence reads from next-generation sequencers and sequence traces from the conventional capillary sequencers respectively (14).

Despite holding the same sequence data, the primary sequence depositories have their own data submission and search interfaces. GenBank accepts either Web-based submissions (BANKIT) suitable for a limited number of sequences with simple annotations (15), or e-mail–based submissions (SEQUIN) suitable for both small and larger sequences with extended annotations (16). Files that exceed the e-mail attachment limits can be submitted using Sequin-MacroSend (17). There are other specific methods for submission of large batches of sequences including EST, sequence tagged sites, and genome survey sequences and sequence reads generated from capillary (trace archive) sequencing machines or one of the next-generation sequencers (sequence read archive) (18). EMBL-Bank uses the online submission portal WEBIN for both individual and bulk sequence submissions (19). Users are required to choose from a selection of templates when preparing their submission. The DNA Data Bank of Japan uses the online submission portal SAKURA for the submission of both short and individual sequences and a mass submission system for the submissions of large files and multiple sequences (20). Submission systems of each of these 3 primary sequence depositories offer intuitive interfaces and the advantage of resuming partial submissions.

Secondary databases differ from primary sequence repositories because the sequences in these databases are curated and nonredundant. The National Center for Biotechnology Information (NCBI) offers a large collection of secondary databases in various sections that can be accessed through the Entrez retrieval system (21). The RefSeq database is an important component of the Entrez system and is composed of sequences of genomic regions, transcripts, and proteins, which are identifiable through entry-specific accession numbers (22). RefSeq is the first landing point for any user looking for curated sequences in genomes, transcriptomes, and proteomes. The database Entrez-Gene (23) offers genomic-centric information such as genomic location and gene products and their attributes and phenotypes, whereas Entrez-Genome (21) contains chromosomal sequences and maps for all completely sequenced genomes. Similarly, the European Bioinformatics Institute (EBI) maintains a large collection of derived databases and analysis programs. Most importantly, the databases Ensembl (24) and Ensembl-Genomes (25) provide access to gene, transcript, and protein sequences and to whole genomes of vertebrate and nonvertebrate organisms, respectively. Map viewers have been integrated in these databases and provide positional information for various genomic features. Users may benefit from using both Entrez and Ensembl because

<table>
<thead>
<tr>
<th>Competency</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics</td>
<td>Studies of genomes and functional and regulatory elements</td>
</tr>
<tr>
<td>Genetic variation</td>
<td>Studies of genome variations</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>Studies of hereditary marks in chromatin (histones, DNA)</td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>Studies of transcripts, including noncoding RNA and micro RNA</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Studies of proteins, including their structure</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>Studies of metabolites in cells, tissues, and body fluids</td>
</tr>
<tr>
<td>Systems biology</td>
<td>Holistic analysis of the cellular biochemical interaction networks</td>
</tr>
</tbody>
</table>

Table 1. Nutrigenomics core competencies and their relevance.
Table 2. Databases in the major omics fields that contribute to nutrigenomics investigations

<table>
<thead>
<tr>
<th>Database</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomics &amp; Epigenomics</td>
<td></td>
</tr>
<tr>
<td>EMBL-Bank</td>
<td><a href="http://www.ebi.ac.uk/emb/">www.ebi.ac.uk/emb/</a></td>
</tr>
<tr>
<td>DDBJ</td>
<td><a href="http://www.ddbj.nig.ac.jp/">www.ddbj.nig.ac.jp/</a></td>
</tr>
<tr>
<td>Ensembl</td>
<td>wwwensembl.org/</td>
</tr>
<tr>
<td>Ensembl-Genomes</td>
<td><a href="http://www.ensemblgenomes.org/">www.ensemblgenomes.org/</a></td>
</tr>
<tr>
<td>UCSC</td>
<td><a href="http://www.ucsc.edu/">www.ucsc.edu/</a></td>
</tr>
<tr>
<td>mirBase</td>
<td><a href="http://www.mirbase.org/">www.mirbase.org/</a></td>
</tr>
<tr>
<td>trRNAdb</td>
<td>trnadb.bioinf.uni-leipzig.de/</td>
</tr>
<tr>
<td>EPD</td>
<td>epd.xit.cch/g/TRED/</td>
</tr>
<tr>
<td>TRED</td>
<td>TRED.map.nig.ac.jp/</td>
</tr>
<tr>
<td>Roadmap</td>
<td><a href="http://www.rodpemapigenomics.org/data/">http://www.rodpemapigenomics.org/data/</a></td>
</tr>
<tr>
<td>Epigenomics</td>
<td></td>
</tr>
<tr>
<td>Human Epigenome Browser</td>
<td><a href="http://epigenomegateway.wustl.edu/">http://epigenomegateway.wustl.edu/</a></td>
</tr>
<tr>
<td>VizHub</td>
<td><a href="http://genomewebbrowser.wustl.edu/">http://genomewebbrowser.wustl.edu/</a></td>
</tr>
<tr>
<td>UCSC Epigenomics Browser</td>
<td><a href="http://www.epigenomeweb.org/">http://www.epigenomeweb.org/</a></td>
</tr>
<tr>
<td>Genetic variants</td>
<td></td>
</tr>
<tr>
<td>Ensembl Variation</td>
<td><a href="http://www.ensembl.org/">www.ensembl.org/</a></td>
</tr>
<tr>
<td>SNPpher</td>
<td>snppher.chip.org/</td>
</tr>
<tr>
<td>CGAP-GAI</td>
<td>ga'i.ncbi.nlm.nih.gov/cgap-gai/</td>
</tr>
<tr>
<td>HuGE Navigator</td>
<td>hugenavigator.net/HuGENavigator/home.do</td>
</tr>
<tr>
<td>dbGaP</td>
<td>vcb.ncbi.nlm.nih.gov/dbgap</td>
</tr>
<tr>
<td>GIWAS central</td>
<td><a href="http://www.giwascentral.org/">www.giwascentral.org/</a></td>
</tr>
<tr>
<td>1000 Genomes</td>
<td><a href="http://www.1000genomes.org/">www.1000genomes.org/</a></td>
</tr>
<tr>
<td>Project</td>
<td></td>
</tr>
<tr>
<td>Transcriptomics</td>
<td></td>
</tr>
<tr>
<td>ArrayExpress</td>
<td><a href="http://www.ebi.ac.uk/arrayexpress/">www.ebi.ac.uk/arrayexpress/</a></td>
</tr>
<tr>
<td>mESAdb</td>
<td>konulab.fen.bilkent.edu.tr/mira/</td>
</tr>
<tr>
<td>Proteomics</td>
<td></td>
</tr>
<tr>
<td>UniProt</td>
<td><a href="http://www.uniprot.org/">www.uniprot.org/</a></td>
</tr>
<tr>
<td>PRIDE</td>
<td><a href="http://www.ebi.ac.uk/pride/">www.ebi.ac.uk/pride/</a></td>
</tr>
<tr>
<td>PeptideAtlas</td>
<td><a href="http://www.peptideatlas.org/">www.peptideatlas.org/</a></td>
</tr>
<tr>
<td>GPMDB</td>
<td>gpmdb.theepm.org/</td>
</tr>
<tr>
<td>RCSB-PDB</td>
<td><a href="http://www.rcsb.org/">www.rcsb.org/</a></td>
</tr>
<tr>
<td>STRING</td>
<td><a href="http://string-db.org/">http://string-db.org/</a></td>
</tr>
<tr>
<td>Metabolomics</td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td><a href="https://www.ebi.ac.uk/chemb/">https://www.ebi.ac.uk/chemb/</a></td>
</tr>
<tr>
<td>CheBI</td>
<td><a href="http://www.ebi.ac.uk/chebi/">www.ebi.ac.uk/chebi/</a></td>
</tr>
<tr>
<td>KEGG Compound</td>
<td><a href="http://www.genome.jp/kegg/compound">www.genome.jp/kegg/compound</a></td>
</tr>
<tr>
<td>ChemSpider</td>
<td><a href="http://www.chemspider.com/">www.chemspider.com/</a></td>
</tr>
<tr>
<td>HMDB</td>
<td><a href="http://www.hmdb.ca/">www.hmdb.ca/</a></td>
</tr>
<tr>
<td>METLIN</td>
<td>metlin.scripps.edu/</td>
</tr>
<tr>
<td>MassBank</td>
<td><a href="http://www.massbank.net/?lang=en">www.massbank.net/?lang=en</a></td>
</tr>
<tr>
<td>NIST Chemistry</td>
<td>webbook.nist.gov/chemistry/</td>
</tr>
<tr>
<td>WebBook</td>
<td></td>
</tr>
<tr>
<td>KNAPSACK</td>
<td><a href="http://www.metabolome.jp/software/">www.metabolome.jp/software/</a></td>
</tr>
<tr>
<td>BMRB</td>
<td><a href="http://www.bmrw.wisc.edu/metabolomics/">www.bmrw.wisc.edu/metabolomics/</a></td>
</tr>
</tbody>
</table>

(Continued)

both databases offer unique perspectives. As an example, the human holocarboxylase synthetase gene (Entrez-Gene ID: 3141, Ensembl ID: ENSG00000159267) has unique transcripts in both Entrez-Gene and Ensembl. Ensembl-Gene is unique by providing the corresponding Refseq alignments and a Blast/primer search for the displayed sequence range. Ensembl-Gene is unique by providing convenient tabular information for all the associated transcripts, and its browser has information for many additional tracks related to variation and regulatory features. Note that some research centers maintain organism-specific databases, e.g., for some plant and animal model organisms (26–29) and specialized databases may focus on features such as micro RNA (30), transfer RNA (31), gene promoters (32), and other regulatory elements (33).

**Genome variations**

The phenotype of an organism is a result of complex interaction between genotype and the environment. This interplay between genes and the environment is further complicated by the individual genomic variations. The Human Genome Variation Society, which funds the discovery, characterization, documentation, and dissemination of genome variation information, maintains a categorized list of the variation databases (34). SNP are the most abundant form of genetic variation observed among individuals. The analysis of the roles of SNP in disease risk attracted considerable attention after 2 large-scale initiatives (the SNP Consortium and the Human Genome Project) generated large SNP datasets (35) (Table 2). There have been many reports in the past decade linking nutrition, genetic variation, and disease risk, as illustrated using the following examples. MTHFR catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, which is a key step in homocysteine metabolism. Two SNP in the MTHFR gene have been characterized at the biochemical level. The C677T variant impairs the stability of the protein, whereas the A1298C variant results in decreased enzyme activity without affecting enzyme stability (36). Individuals who are T677 homozygous have higher cardiovascular disease risk than C677 homozygous individuals, and the risk can be reduced to that of C677 homozygous individuals by folate supplementation (37). Heterozygotes have elevated homocysteine levels, but a causal link with cardiovascular disease risk is uncertain. Holocarboxylase synthetase activates biotin-dependent carboxylases by covalently attaching biotin to
the apocarboxylase. The SNP A2096G in the coding region of holocarboxylase synthetase gene decreases biotin-binding affinity and thus reduces enzyme activity. Supplementation with biotin can restore activity to wild-type levels (38) at least in vitro.

Currently there are >30 million human reference SNP in the SNP Database (dbSNP), which is the largest repository of SNP. dbSNP also contains less frequent types of variations such as multibase deletions or insertions, including those of retroposable elements and microsatellites (39). Distinct phenotypes have only been identified for a minority of the known SNP. Some rare SNP might have escaped detection. The dbSNP can be searched using keywords, or entries can be retrieved through unique accession numbers. Approximately half of the SNP in dbSNP have not been validated. Accordingly, entries must be interpreted with caution (40).

Numerous other databases are viable alternatives to dbSNP in studies of genome variation (Table 2). The Ensembl variation database is one such option. Many of the Ensembl data are imported from the dbSNP and other databases, and the data are linked with consequence types revealing effects on the final protein product (41). In addition, the University of California, Santa Cruz (UCSC) genome browser maintains its own collection of genomic variants (42). The Online Mendelian Inheritance in Man is one of the earliest established resources linking human diseases with allelic variants including SNP (43). Online Mendelian Inheritance in Man is indispensable for medical researchers as a comprehensive source of information for all the known disease-associated genetic variants. Other databases focus on SNP in specific organisms, specific diseases, or functional metadata. Many of these databases also have tools for experimental analysis, such as designing PCR primers. SNPper is an example for an organism (human)–specific SNP database. SNPper is derived from the dbSNP and the UCSC Human Genome Browser (44). Cancer Genome Annotation Project–Genetic Annotation Initiative’s genetic variation resource is an example of a disease (cancer)–specific database (45). In most of these databases, SNP can be visualized in the context of genes, transcripts and regulatory features through the use of integrated genome browsers.

As a result of the decreasing costs of high-throughput sequencing, genome-wide association studies (GWAS) have become a common strategy for identifying the genetic basis of disease susceptibility. GWAS involves surveying SNP in samples from a large number of healthy and afflicted individuals to identify variations that associate with disease phenotypes (46). The Database of Genotypes and Phenotypes at the NCBI is one of the largest GWAS repositories (47). The summary data are freely accessible, whereas individual level data are restricted to approved users. GWAS central is an alternative portal for accessing summary level information and has the added advantage of providing an integrated viewer to visualize the location of markers at the chromosomal and gene level (48). The Human Genome Epidemiology navigator collects and classifies genetic variant information from the literature and is an example for a database providing information about meta analyses and GWAS (49). Human Genome Epidemiology navigator can be searched using gene names (Genopedia) or disease names (Phenopedia). The value of SNP as markers for disease risk is undisputed. However, despite recent advances in genotyping protocols, it is still not feasible for single investigators to interrogate a substantial fraction of the SNP in the human genome. In this regard, haplotype analysis has gained traction. Haplotypes are the specific patterns of SNP over short stretches of genome that tend to inherit together. Genotyping using the haplotype information for certain “tagged SNP” is more economical and has little information loss (50). The Ensembl variation database includes a list of “tagged variants” that have been identified as having high linkage disequilibrium with other proximal variants. These are useful for haplotype analysis. Haplotypes are catalogued in the HapMap Project (51) and in the 1000 Genomes Project (4). The variant and haplotype information from these databases can be combined in individual disease association studies to obtain genotype information beyond what is genotyped directly, which helps the researcher to precisely locate disease-associated regions.

**Epigenomics**

Apart from the genome, the organism’s phenotype is influenced by its epigenome, which is the full complement of epigenetic marks including DNA methylation and histone modification. Micro RNA and other noncoding RNA are also being classified as epigenetic factors due to the accumulating evidence of their role in mediating epigenetic mechanisms like chromatin remodeling and transcriptional and posttranscriptional regulation (52) (see further discussion in the Transcriptomics section). Epigenetic factors can control gene expression (53), DNA replication (54), DNA repair (55), and DNA recombination (56). Epigenetic factors exhibit a dynamic profile depending on the environment status and developmental status of the organism. Micronutrients and bioactive food compounds play essential roles in creating epigenetic marks [see Zempleni et al. (57) for a review]. Typically, epigenetic marks are mapped using chromatin immunoprecipitation (ChIP) assays, in combination with DNA microarrays (ChIP-chip) or high-throughput sequencing (ChIP-seq) (58). MethDB is one of the foremost public databases dedicated to storing DNA methylation data (Table 2). The database provides graphical and tabular representation of methylation profiles, including experimental details and sample phenotype characteristics; the database can be searched by species, sex, tissue, and gene (59). Because alterations in DNA methylation are implicated in tumorigenesis (60), cancer-specific methylation databases such as PubMeth have been developed. PubMeth was created by literature mining and manual curation; the database can be searched by gene and by type of cancer (61).

Researchers with an interest in histones may use the Human Histone Modification Database, which currently holds data for 43 types of histone modifications in humans; these
were produced by ChIP-chip, ChIP-seq, and qChIP experiments. The database is searchable by type of histone modification, gene, chromosomal location, functional category, and type of cancer; individual datasets can be viewed in the integrated genome browser (62). Additional databases provide extensive information on histones (63), histone-modifying proteins (64), other chromatin-associated proteins (65), and chromatin-remodeling proteins (66).

The Encyclopedia of DNA Elements Project also generated valuable information for epigenetics research and offers a catalogue of functional DNA elements including genes, transcripts, cis-regulatory elements including promoters, enhancers, insulators, silencers, transcription factor binding sites, DNA methylation sites, and histone modification sites (67). Encyclopedia of DNA Elements data can be easily viewed using the publicly available genome browsers, like the UCSC genome browser (42), NCBI workbench (68), GBrowse (69), and the ENSEMBL browser (70).

The NIH Roadmap Epigenomics Mapping Consortium was launched in 2008 to provide a publicly accessible resource of reference epigenomic maps in stem cells and primary ex vivo tissues from different individuals, mapping DNA methylation, histone modifications, and related chromatin features, with the objective to aid basic and disease-oriented research (71). The data generated by the consortium can be browsed from the project Web site using a matrix, in which rows correspond to tissue/organ types and columns correspond to epigenetic variables. The data can also be browsed using the visual data browser for specific stem cells and fetal and adult tissues. The consortium’s data can be visualized with the dedicated Human Epigenome Browser maintained by Washington University (72). The data tracks are uniquely represented as heat maps with color gradients denoting signal strength. The data can be selected based on cell type, assay type, epigenetic mark, phenotype, and data source. The browser also facilitates standard statistical analysis on the tracks including pairwise comparison, hypothesis testing, and correlation. The data can also be accessed using the human epigenomic atlas, the NCBI epigenomics hub, and the UCSC browser mirror for epigenomics data (Table 2). Each of these resources have unique features with respect to data representation, browsing method, data download formats, and data upload capability; each resource offers additional tools for viewing and analyzing the data (73).

**Transcriptomics**

Historically, the quantification of mRNA has been at the forefront of transcriptomics. More recently small RNA such as micro RNA and noncoding RNA have attracted considerable attention (74). Microarrays and next-generation sequencers are the primary analytical technologies in transcriptomics research; users can select from various platforms. The microarray technology is mature and well established. However, there are certain difficulties with data analysis and the reproducibility of the results, especially in the context of nutritional research due to the complex nature of relationships between nutrients and the target genes. Transcriptional profiling using microarrays has been used to identify cellular targets for many macronutrients and micronutrients and also to characterize gene expression differences under different nutritional conditions (75).

In the recent past, microarrays played a major role in expression profiling, and they continue to be used as a lower cost approach to gene expression analysis. Their use is limited because only known genes can be studied. Tiling arrays use contiguous stretches of genomic regions that cover both known and unknown genes; however, their use depends on the availability of a reference DNA sequence. Other contemporary sequencing methodologies, EST analysis, and serial analysis of gene expression do not have this limitation, but limited sensitivity of transcript identification and high costs have proven prohibitive for many laboratories. Currently, the transcriptomics investigations increasingly use high-throughput RNA-seq technology. In addition to providing quantitative data and a greater dynamic range of detection, RNA-seq has the ability to detect sequence variants and splicing events without bacterial cloning (76,77).

The Gene Expression Omnibus (GEO) at NCBI is a central depository hosting both sequence-based and microarray-based gene expression data (78) (Table 2). The data can be quickly located either by searching the GEO DataSets database or by searching for specific gene expression profiles in the GEO Profiles database. This database offers a number of tools for statistical analysis of treatment-responsive genes, for cluster analyses, and for mapping the results using NCBI-BioSystems records in pathway analyses. The GEO2R tool is useful for identifying genes that are differentially expressed between ≥2 groups. The ArrayExpress archive (79) at EBI is another large public repository for microarray data from all platforms. The database can be queried for experiment, organism, array type, and author. The Gene Expression Atlas (80) is associated with the ArrayExpress archive and contains a database of summary statistics based on meta-analysis from a curated subset of expression data. The Atlas is useful for retrieving gene expression patterns in various organisms in different tissues and under different environmental, disease, or developmental states. With the recent discovery of role for noncoding RNA in biological systems, studies are now emerging that focus on profiling the small RNA transcriptome (81); this has led to the establishment of dedicated databases (82–84).

Several steps in the analysis and interpretation of the microarray data including data normalization and supervised and nonsupervised methods of differential gene expression analysis can be performed by well-established open source software like Bioconductor/R TM4 software suite (85) and by other free tools (86). Next-generation platforms for RNA sequencing generate terabytes of data. These platforms use specific methods for quality assessment, alignment, assembly, and further processing (87), which are specific for noncoding RNA (88). Bioconductor/R can be used to perform data analysis (85), but alternative software suites and stand-alone programs are also available (89).
**Proteomics**

Proteomics focuses on the identification and the quantification of all cellular and extracellular proteins. Initially, 2-dimensional gel electrophoresis assays were frequently used for analysis of protein profiles and resulted in the establishment of many databases for these data (90). Later, protein microarrays based on the principle of immunoassays in which affinity reagents (e.g., antibodies) are spotted onto the arrays at high density were used to identify and quantify proteins (91). These technologies have important limitations; for example, 2-dimensional electrophoresis is less sensitive and less reproducible than protein arrays. Two-dimensional electrophoresis is not useful in comparison of profiles from different species. Protein arrays are limited by the requirement for a large number of high-quality antibodies. In this context, MS methods, especially matrix-assisted laser desorption ionization time of flight MS and electrospray ionization MS, have emerged as the most useful proteomics methods (92).

Database resources in proteomics are not as comprehensive as those in genomics and transcriptomics, and proteomic databases are heterogeneous with respect to content, primarily because of the high degree of complexity in the methods used for proteomics research. The publicly available proteomics databases complement each other to serve the varied needs in proteomics research (93) (Table 2). The PRIDE database at EBI is a prominent centralized repository for MS proteomics data that holds spectral data and data for peptide and protein identification. Public data as well as collaborative data can be accessed using a simple search based on protein accession numbers or using the advanced search, which uses peptide sequence or accession numbers from other databases. The results can be viewed online or downloaded in XML (extended markup language) files. Specific experiment sets can be compared creating visual Venn diagrams. The PRIDE-Inspector tool can be installed locally and used to interact with the database and view the results locally (94). The PeptideAtlas database at the Institute for Systems Biology is highly useful in targeted proteomics work because it provides readily accessible peptide information of high confidence extracted from existing MS spectral data. PeptideAtlas has an extensive architecture and compiles its data into several builds; in these, the raw data are processed through a series of validation, peptide identification, and genome mapping steps before making the entry accessible for users. The database can be searched against various builds using either the peptide/protein name or the peptide sequence; other useful features include searching using protein lists and searching for specific pathways (95). The Global Proteome Machine database (GPMDB) is the largest curated public proteome repository. GPMDB contains MS/MS spectra along with protein identifications (96). The database provides the X! Hunter and X! Tandem search engines, which can be used to compare user input spectra with the consensus spectra. These search engines can also identify proteins for the selected sample sources including human proteins. Users may deposit spectra in the GPMDB. Identified proteins can be viewed either in gene, protein sequence, or ontology representations.

The peptide sequences from the mass spectral analysis are usually mapped to protein databases for protein identification. The most comprehensive database in this regard is the UniProtKB (97). UniProtKB has 2 sections: TrEMBL, which contains automatically translated and annotated EMBL sequences, and Swiss-Prot, which contains nonredundant sequences that are manually annotated and reviewed. The databases can be searched by accession numbers, keywords, or protein sequence. The result pages have detailed information on function of the protein with links to specialized databases. Three-dimensional structural information on proteins can be accessed from Research Collaboratory for Structural Bioinformatics PDB database, which is a repository for experimentally determined 3-dimensional structures of proteins (98).

Finally, information on protein–protein interactions is necessary to develop pathways and molecular interaction networks. The Search Tool for the Retrieval of Interacting Genes/Proteins database at EBI contains known and predicted protein–protein interaction data from various organisms (99). Database entries are based on high-throughput experiments, literature mining, and sequence homology. The interactions are provided with confidence scores and can be viewed as a network or in other formats such as coexpressing proteins, neighborhood genes, fusion partners, pathway database links, links to specific experimental details, or PubMed citations. A large number of proteomics programs can be found at the ExPaSy bioinformatics resource portal (100).

**Metabolomics**

Metabolomics is the study of all small molecule metabolites in living systems (i.e., cells and tissues) and their biological fluids (e.g., blood and urine) and deals with their identification and quantification (101). The metabolome is the most dynamic component of living systems with respect to changes in composition and relative proportions over time and in response to various biotic and abiotic factors. The study of the metabolome has a wide range of applications including studies of pharmacogenetics (102) and nutritional intervention (103). The metabolome is complex and diverse; the constantly changing metabolite flux poses unique challenges. Investigators may choose to focus on single metabolites (targeted analysis), groups of metabolites belonging to a specific class or pathway (profiling), or the unbiased identification and quantification of all metabolites (fingerprinting and footprinting) (104). For general information about metabolomics, relevant technologies, databases and data analysis tools, the user is referred to the metabolomics links at Scripps Institute (105).

The 2 most widely used analytical techniques in metabolomics are MS and NMR spectroscopy. MS techniques are indispensable for identification and characterization of unknown metabolites. Coupled with various chromatographic/electrophoretic separation methods, they offer
excellent sensitivity. However, MS approaches are limited to ionizable metabolites (104). Although sensitivity is lower and dynamic range is smaller for individual identification of metabolites than MS, NMR has advantages. NMR requires little or no sample preparation and separation. NMR yields quantitative data, which are highly suitable for fingerprinting (106). Although less sensitive than MS and NMR, Fourier transform infrared spectroscopy is increasingly used as a rapid and nondestructive technique in fingerprinting.

Based on the type of primary information, metabolomic databases can be grouped into those that harbor the following a) information about the physicochemical properties and biological annotations of metabolites; b) experimental raw data, in the form of spectra coming from various analytical platforms (mostly MS and NMR); and c) information regarding the pathways that connect metabolites (107). Some databases are hybrids of these 3 forms (Table 2).

**Databases with information on metabolites**

The databases in the first category are dictionaries of small molecules, providing structure and biological activity information for endogenous metabolites and xenobiotics. In this category, PubChem is probably the most comprehensive small molecule database, offering physicochemical information, bioassay information, and a literature mining tool (108). ChEMBL (109) and ChEBI (110) at EMBL-EBI focus on pharmacologically active molecules and chemicals of biological interest, respectively. KEGG Compound in Japan also is useful source of information (111). The Royal Society of Chemistry, London, maintains ChemSpider (112), which contains information on chemicals including related metabolites; ChemSpider incorporates data from >150 sources and provides binding predictions for target/receptor proteins based on LASSO scores in an unclustered format that is useful for molecular interaction studies. In addition, databases are available for metabolites and other small molecules from distinct compound classes such as carbohydrates (113), lipids (114), drugs (115), environmental chemicals (116), and metabolites in specific organs (117). Many of these databases fail to provide disease-relevant information such as effective concentrations, sources, and methods used for experimental identification and quantification. To serve this purpose, the Human Metabolome Database (HMDB) was developed as a part of the Human Metabolome Project. HMDB aims to catalogue all metabolites present in body fluids in significant concentrations and to validate the information for known metabolites (118). HMDB is the largest central database dedicated to the study of the human metabolome. An important feature of the HMDB is the incorporation of experimental data from specific analytical platforms, thus allowing an easy search and comparison of metabolites detected using various experimental techniques.

**Databases holding raw data on metabolite detection**

Databases in this category store the experimental data in the form of spectra or chromatograms from different analytical platforms and are analogous to the high-throughput genomics databases. They provide the reference data useful for the identification and quantification of unknown compounds as well as metabolic profiles for diagnostic purposes. Major databases providing reference mass spectral data include METLIN (119), MassBank (120), NIST Chemistry WebBook (121) and KNAPSACK (122). These databases can be searched by accession numbers, compound names, chemical formula, substructure, mass, and mass/charge ratios. METLIN has annotations for >42,000 metabolites and tandem mass spectral data for many of them. MassBank is a distributed metabolite database composed of mass spectra originating in several collaborating institutions. NIST contains electron impact MS and MS/MS spectral libraries, gas chromatography data libraries, and a Web service for converting between gas chromatography/MS data formats from various platforms. KNAPSACK contains species-specific data to facilitate comparative studies of mass spectra in relation to organism and taxonomic hierarchy. NMR is an indispensable tool for structural identification of completely novel compounds (123). NMR reference spectral data are essential components of NMR-based metabolomics. Biological Magnetic Resonance Data Bank (124) is a freely accessible, curated database of original reference NMR data sets. Other notable NMR spectral databases include NMRshiftDB2 (125) and ChemSpider (112). PRIME (126) is a Web-based service that harbors NMR and MS reference spectra and several tools for an integrated analysis of metabolomics and transcriptomics data. The Madison Metabolomics Consortium Database (127) presents the user with additional information on the metabolite in ~50 separate data fields.

**Databases storing metabolic pathway information**

Databases in this category store pathway information. Such data are useful to predict and compare metabolic pathways in newly sequenced organisms and also aid in metabolic engineering in the context of disease prevention and drug/dietary intervention studies. Kyoto Encyclopedia of Genes and Genomes (KEGG) (128) is among the oldest and best established pathway databases. KEGG contains manually compiled pathway maps with metabolite and reaction information from published literature. KEGG offers sequence similarity and chemical structure similarity search features; displays may be color coded. Reactome (129) is another curated and peer-reviewed pathway database with cross-references to other databases containing information about nucleotides, proteins, metabolites, and molecule interactions. Reactome has an emphasis on chemical reactions, and the networks of molecules participating in the reactions are grouped into pathways. In addition to the human pathway data, Reactome provides inferred pathway information for 20 other model organism/species and provides tools for comparative analysis of pathways. BioCyc (130) is a large repository of ~1690 organism-specific metabolic pathway databases. BioCyc is organized into 3 tiers based on the level of data curation. The top-tier databases are intensively curated;
these include MetaCyc, which is a large collection of metabolic pathways from multiple organisms; HumanCyc for human pathways; EcoCyc for Escherichia coli pathways; AraCyc for Arabidopsis pathways; and YeastCyc for yeast pathways. The other 2 tiers provide information for computationally derived data with moderate (tier 2) or no curation (tier 3). The BioCyc Web site also provides several tools for metabolic pathway analysis including network analysis, comparative pathway analysis, and data visualization. The Small Molecule Pathway Database has >350 small molecule pathways, of which >280 are unique to this database (131). The pathways are depicted in graphical form in their cellular location, and the concentrations of metabolites in pathways are shown. The Small Molecule Pathway Database can be searched by text or by SwissProt, GenBank, or Affymetrix/Agilent microarray identifiers.

**Systems biology**

Systems biology studies the organism/system as a whole and is based on the premise that the information content of the interacting parts of the whole system studied together is more than the sum of its parts, revealing what are called the emergent properties of the system (132). Traditionally, biological research has been conducted using a reductionist approach in which the individual components are investigated in detail to identify cause-and-effect relationships. However, biological systems function as interacting components rather than as isolated components (133). Although systems biology methods had been appreciated since the 1950s (134), its current popularity and wide spread use is the result of the recent developments in “omics” technologies that create the large datasets necessary for a systems approach. Systems biology integrates the various “omics” datasets from genes to metabolites to model interaction networks and to study network function and evolution under normal and perturbed conditions (135). Systems biology can identify disease mechanisms that were previously not apparent; especially those of complex, chronic diseases. Nutrigenomics more than other fields lends itself to systems biology approaches due to the multiple input variables and the large number of diet–molecule interactions; most of these interactions cause only subtle effects that are not directly apparent if investigated in isolation (3). For example, modeling of folate-mediated 1-carbon metabolism pathway and simulation of the impact of genetic and nutritional variation gave novel insights related to the pathway (136). Low folate is inversely associated with serum homocysteine levels, but no associations have been established with DNA methylation marks. Decreased MTHFR activity due to the C677T polymorphism may lower levels of S-adenosylmethionine, 5-methyltetrahydrofolate, and DNA methylation, whereas increasing the levels of S-adenosylhomocysteine, homocysteine, and the rate of purine synthesis. Folate deficiency may increase the effects size in C677T homozygotes.

Some online resources are available for studies that take a systems biology approach. The basic requirement is that the entire genome of the model organism has been sequenced. A systems biology approach usually involves studying the biochemical network, by perturbation of the genetic or the environmental components of the system, in an iterative process of hypothesis generation and testing of the hypothesis to generate a new hypothesis (137). Therefore, the availability of databases that store molecular interactions and pathway information is essential. Because the amount of data that needs to be retrieved is large, platforms such as BioMart have been developed to interact and retrieve data from the various omics databases (138). BioMart was designed to work with any existing or newly developed databases. This benefits the user by relying on only 1 unified data access interface without having to master the advanced Web query interfaces or application programmatic interfaces that are specific to each data source. The power of BioMart comes from integrated querying of different Marts, whose source data locations may be at geographically distant regions, and from its easy integration with other external software. Software like Galaxy provides an integrated data analysis interface. Galaxy (139) offers a seamless approach to data analysis, including basic sequence searches and genome-wide analysis of next-generation sequencing data. Galaxy also offers tools for data storage and management. Thus, Galaxy eliminates the bottleneck at data integration and large-scale data management that can be a limiting factor in systems biology investigations. Gaggle is another tool that can link the capabilities of different software tools operating on varied data into an integrated analysis framework (140). Gaggle is a simple and extensible system; any software can be adapted into its framework with a minimum amount of coding for suitable modifications in the input and output data formats. Molecular interaction networks can be modeled, visualized, and analyzed using the free Cytoscape package (141). Cytoscape has several plug-ins suitable for several uses in systems biology, including advanced analysis and modeling of interaction data, with powerful visual mappings with functional annotations. Additional resources are constantly being developed in this relatively new field of research.

**Conclusion**

Nutrigenomics investigators seek to understand the organization and function of cellular components and characterize the various molecular phenotypes associated with health and disease. These studies are facilitated by omics technologies, which have created unprecedented opportunities. With the cost of sequencing decreasing steadily, the costs of data storage and analysis may prove to be the true bottlenecks in moving the field forward (142). This situation will likely worsen with the increasing amount of high-resolution data arising from complex disease and metagenomic investigations of the role of microbial communities in various ecological niches, including the human intestine (143). A greater understanding of the genome at the level of individual variations could eventually lead to the development of the much anticipated paradigm of personalized nutrition and...
We hope that this review will prepare the ground for nutri-
-genomics investigators to embrace bioinformatics and to

take full advantage of the opportunities in this field.

Acknowledgments

All authors have read and approved the final manuscript.

Literature Cited

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the fu-


2. Collins FS, Morgan M, Patrinos A. The Human Genome Project: les-


3. Mutch DM, Wahl W, Williamson G. Nutrigenomics and nutrige-


4. 1000 Genomes Project Consortium. A map of human genome varia-


5. Pennisi E. Genomics. 1000 Genomes Project gives new map of genetic


6. Botto LD, Yang Q. 5,10-Methylene tetrahydrofolate reductase gene var-


7. Wrone EM, Zehnder JL, Hornberger JM, McCann LM, Coplon NS,

Fortmann SP. An MTHFR variant, homocysteine, and cardiovascular


8. World Health Organization. World Health Assembly: Genomics and

World Health. Fifty Seventh World Health Assembly Resolution.

2004;World Health Assembly 57.13. [cited 2012 July 21]; Available:

pdf.

9. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and ge-


10. Benson DA, Karsh-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. Gen-


cbbl.nlm.nih.gov/LargeDirSubs/dir_submit.cgi.

12. National Center for Biotechnology Information. The GenBank Sub-


cbbl.nlm.nih.gov.iii-server.uarh.edu/books/NBK51157/.

13. Amid C, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I,

tool developments at the European Nucleotide Archive. Nucleic


14. DNA Data Bank of Japan. SAKURA. 2012. [cited 02/20/2012]; Avail-

able from: sakura.ddbj.nig.ac.jp.

15. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K,


resources of the National Center for Biotechnology Information. Nucle-


gov/LargeDirSubs/dir_submit.cgi.

17. National Center for Biotechnology Information. SequinMacroSend.


gov/LargeDirSubs/dir_submit.cgi.

18. National Center for Biotechnology Information. The GenBank Sub-


cbbl.nlm.nih.gov.iii-server.uarh.edu/books/NBK51157/.

19. Amid C, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I,

Faruque N, Gibson R, Goodgame N, Hunter C, et al. Major submis-

sions tool developments at the European Nucleotide Archive. Nucleic


20. DNA Data Bank of Japan. SAKURA. 2012. [cited 02/20/2012]; Avail-

able from: sakura.ddbj.nig.ac.jp.

21. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K,


resources of the National Center for Biotechnology Information. Nucle-


22. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Se-

quences (RefSeq): current status, new features and genome annota-


24. Flicek P, Amode MR, Barrett T, Beat K, Brent S, Carvalho-Silva D,


25. McQuilton P, St Pierre SE, Thurmond J, FlyBase 101

26. Botto LD, Yang Q. 5,10-Methylene tetrahydrofolate reductase gene var-


27. Wrone EM, Zehnder JL, Hornberger JM, McCann LM, Coplon NS,

Fortmann SP. An MTHFR variant, homocysteine, and cardiovascular


28. World Health Organization. World Health Assembly: Genomics and

World Health. Fifty Seventh World Health Assembly Resolution.

2004;World Health Assembly 57.13. [cited 2012 July 21]; Available:

pdf.

29. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and ge-


30. Benson DA, Karsh-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. Gen-


32. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T,

Sugawara H, Ogasawara O, Takagi T, Okubo K, et al. DDBJ progress


33. National Center for Biotechnology Information. Accession Number pre-

fixes: Where are the sequences from? 2012 [cited 2012 02/20/2012]; Avail-


34. Leinonen R, Sugawara H, Shumway M, International Nucleotide Se-

quence Database C. The Sequence Read Archive. Nucleic Acids Res.


tool=genbank.

70. Fernandez-Suarez XM, Schuster MK. Using the Ensembl genome browser to browse genomic sequence data. Current Protoc Bioinformatics. 2007 Jan;Chapter 1:Unit 1 15.


