Building the bridges to bioinformatics in nutrition research[1–3]

Danielle G Lemay, Angela M Zivkovic, and J Bruce German

ABSTRACT

Like other life sciences, nutrition science can benefit enormously from the techniques of bioinformatics. In this article, the steps necessary to enable bioinformatic approaches in nutrition research are outlined, from the short-range goal of immediately making data available in ad hoc author-defined formats to the longer range goals of full standardization of nutrition experiments and migration of all experimental data into databases. Several examples of what will be possible for nutrition researchers in this new paradigm are described. Ultimately, nutrition data can be continually recycled to reinvestigate existing hypotheses and to generate new hypotheses that would not have been conceivable at the time of the original experiments. The standardization of experimental designs and the conversion of nutrition data into a machine-readable format will bring about a renaissance in nutrition research, accelerating the ability of investigators to discover the implications of nonessential nutrients and food components, and enable the study of complex metabolic interactions in human health and disease. *Am J Clin Nutr* 2007;86:1261–9.

KEY WORDS Nutrition, bioinformatics, standardization, nutrition standards, biomarkers, bioprofiles, biomarker discovery, nutrigenomics, systems biology, informatics, meta-analysis, computational biology, applied bioinformatics, databases, ontologies, genomics

INTRODUCTION

The greatest advances in life sciences during the past 20 years have arguably been made possible largely by the technologies of computing that are now brought to practice in scientific fields, from analytic chemistry to mathematical simulations. Nutrition, being a highly integrative science that draws from many disciplines, likewise has the potential to benefit enormously from the application of these computational techniques. An obvious prerequisite for the application of bioinformatic techniques in nutrition is the accessibility of nutrition data in machine-readable formats. This article describes the action items that, if pursued, will enable data integration and analysis across all nutrition studies. These goals, along with the roles of various entities in achieving them as summarized in **Figure 1**, are discussed as are concrete examples of what will be possible for nutrition researchers in this new paradigm.

FIRST EXAMPLE OF APPLIED BIOINFORMATICS IN NUTRITION RESEARCH

To understand the mechanisms of nutrient action, nutrition researchers necessarily use a reductionist strategy, breaking the problem down to cells, proteins, genes, etc, and then reintegrating the knowledge gained with higher levels of abstraction to arrive at human body–level theories of nutrient effects. Thus, nutrition researchers regularly generate and interpret data at the molecular level. The development of a comprehensive, predictive understanding of metabolism requires that nutrients and metabolites be explored within the context of their associated regulatory mechanisms. Peroxisome-proliferator activated receptors (PPARs) provide one such example of molecules that directly link nutrient intake to organism response. PPARs are transcription factors that detect various metabolites, including fatty acids and fatty acid derivatives, at the cellular level and then, in turn, launch a specific metabolic program by regulating the expression of a variety of target genes.

As an important step toward a complete mechanistic understanding of PPARs, a recent bioinformatics study was conducted to predict PPAR gene targets on a genome-wide basis (1). This study effectively provides a first library of nutrient-sensitive genes and a first demonstration of how databases and software can be integrated to investigate nutritionally relevant biological questions, such as "Which genes are directly regulated by PPARs and, thus, by fatty acids and fatty acid derivatives? What are the biological functions of these fatty acid–responsive genes? What other transcription factors regulate these fatty acid–responsive genes"? A simplified flowchart in **Figure 2** illustrates how databases and software were integrated to answer these questions. With the exception of the professional TRANSFAC database, all other databases and software tools were publicly available. Development of some customized software was required because there is no commercial biological analysis software package that integrates every type of data and performs the desired analysis tasks. However, the success of this project demonstrates that if the necessary databases are appropriately formatted, annotated,

[1] From the Department of Food Science and Technology, University of California, Davis, CA (DGL, AMZ, and JBG), and Nestlé Research Centre, Lausanne, Switzerland (JBG).

[2] Supported in part by the National Institute of Environmental Health Sciences (NIEHS) grant R37 ES02710, the NIEHS Superfund Basic Research Program P42 ES04699, the University of California Davis Center for Children's Environmental Health, NIEHS grant P01 ES11269, and the California Dairy Research Foundation.

[3] Reprints not available. Address correspondence to JB German, Department of Food Science and Technology, 1 Shields Avenue, University of California, Davis, CA 95616. E-mail: jbgerman@ucdavis.edu.

Received February 21, 2007.

Accepted for publication March 27, 2007.

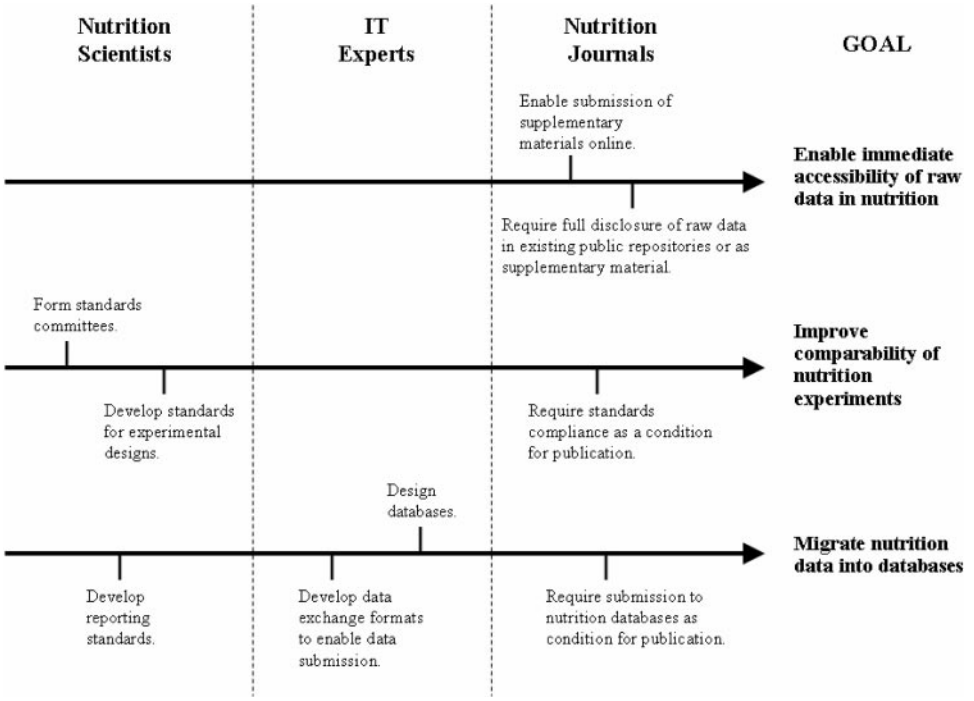

FIGURE 1. Achievement of 3 major goals can enable bioinformatics in nutrition. Nutrition scientists, computer programmers, and nutrition journals have separate milestones to reach these goals. Dependencies between milestones are depicted along the arrows from left to right. IT, information technology.

and publicly accessible, a single investigator can make a significant scientific contribution to nutrition research using current bioinformatic techniques.

Unfortunately, despite its potential, bioinformatics has not reached the mainstream of nutrition research. Beyond the analysis of microarray data in an individual research project or the occasional use of BLAST to locate individual genes or proteins, the full power of bioinformatics in nutrition research remains largely underutilized, principally because of a paucity of comparable electronic nutrition data. The data that enabled the study diagrammed in Figure 2 arose from pharmacology and genomic efforts, not nutrition. Until nutrition data are collected in centralized, publicly accessible databases, the use of bioinformatics in nutrition research will be limited. The following sections describe specifically what steps need to be taken (Figure 1) for nutrition researchers to enable advanced bioinformatics research in nutrition.

ROLES OF MAJOR NUTRITION JOURNALS

Major scientific journals in molecular biology played a crucial role in enabling the PPAR study by requiring disclosure of data sets in public repositories. After the Minimum Information About a Microarray Experiment (MIAME) standard (10) was introduced, adoption of the standard by the scientific community was effectively enforced through modifications of journal policies. In October 2002, the *Nature* group, the *Lancet*, *Cell*, and the *EMBO Journal* adopted the MIAME specification as a requirement for the publication of microarray experiments. Most journals now require that authors of manuscripts that describe microarray data submit their data to a public database such as the NCBI gene expression omnibus (GEO) hybridization array data repository (11) and reference the GEO accession number in the manuscript.

In addition to microarray experiments, journals require authors to provide accession numbers to databases to identify molecular structures in their manuscripts. For example, DNA sequences are referenced by National Center for Biotechnology Information (NCBI) accession numbers such as GenBank (12) or Entrez Gene IDs (13). Protein sequences are identified by accession numbers such as UniProt (14). Protein structures are referenced in the Protein Data Bank at Research Collaboratory for Structural Bioinformatics (15).

For other data for which public repositories do not yet exist, scientists can make their data available to other scientists by submitting it as supplementary material for the online version of their manuscript. Most life science journals outside of nutrition do have such online provisions and are supportive of the recommendations of the National Academies regarding data sharing (16). In addition to providing the means to store and display raw data sets, some journals explicitly require such disclosure for publication.

Major nutrition journals have the power to profoundly accelerate progress in the field of nutrition through journal policy. Publication requirements can be revised to require that data sets be fully disclosed, either by repository to a public database using standardized formats when such exists or by submission to the journal as online supplementary materials. Although journals may incur some cost to provide online storage of raw data sets, journals are rewarded with higher impact factors for articles that are cited not only for their conclusions, but also for their data. In nutrition, data are often presented in the form of statistical summaries to support the specific conclusions of that manuscript. If raw data sets were available, additional hypotheses could be investigated by other scientists. Finally, because standards unique to nutrition research are introduced to facilitate interstudy analyses, the major journals in nutrition will have the unique

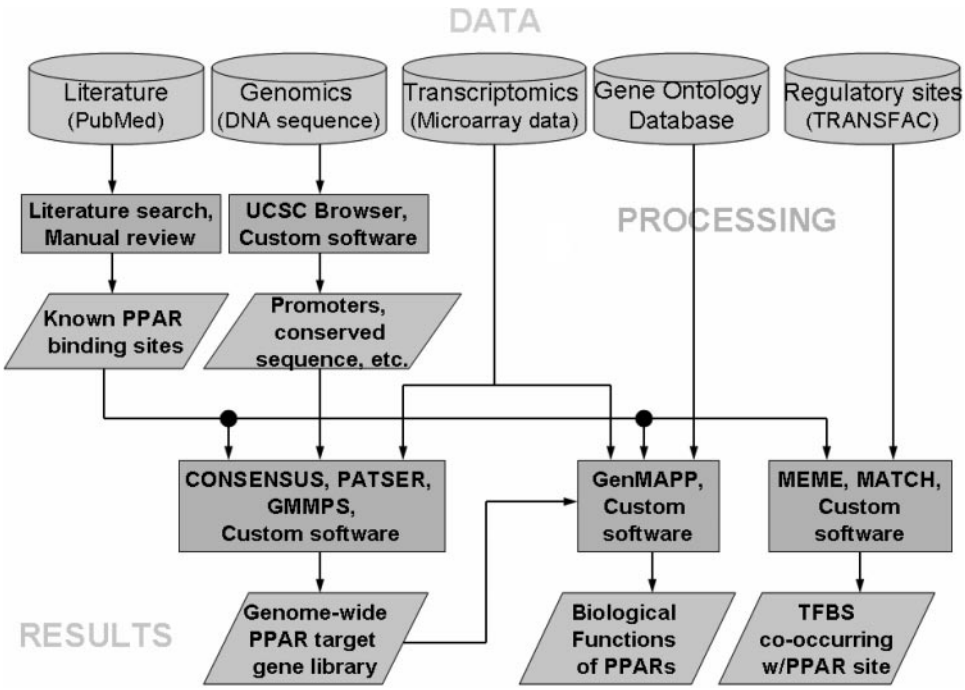

FIGURE 2. Integration of databases and software to predict genes regulated by peroxisome-proliferator activated receptors (PPARs). Literature in the PubMed database was searched for manuscripts containing experimental evidence for DNA binding sites of PPARs. These known PPAR binding sites were used to build probability matrices with different probabilistic assumptions with the use of the CONSENSUS (2) and GMMPS (3) programs. The University of California at Santa Cruz (UCSC) Table Browser (4) and some custom programs were used to extract relevant genomic information—all known human genes, regions of DNA upstream from their transcription start site, conserved elements within these upstream regions, and homologous genes in the mouse and rat genomes. The probability matrices were scored against DNA sequence upstream from known PPAR target genes and randomly selected genes in the genome using the PATSER program (2) and custom software. Techniques that minimized the number of false-negative and false-positive results in the detection of PPAR binding sites were then evaluated by using data from microarray studies. The technique best able to differentiate between regulated and nonregulated genes in microarray studies of PPAR agonists was then applied to the whole genome to identify putative PPAR target genes on a genome-wide basis. The 3 sets of genes—known PPAR targets, those regulated by PPARs in microarray studies, and the predicted genome-wide PPAR targets—were analyzed by using the Gene Ontology database (5) with GenMAPP (6), a gene ontology analysis tool, along with custom software to determine the biological functions represented by each group. Additionally, the TRANSFAC database (7) was used along with the MATCH program (8) on the promoter regions of known PPAR target genes to determine whether any other known transcription factor binding sites (TFBS) were more enriched among these genes compared with random genes. A similar strategy was used with a motif discovery tool called MEME (9) and custom software to discover novel motifs that are more prevalent among known PPAR target genes. "Custom software" refers to the ≥100 Perl and shell scripts that were developed to integrate the databases and tools and to conduct computational experiments.

opportunity to lead changes in the field by requiring compliance with such standards as a prerequisite for publication in their journal.

DEVELOPMENT OF STANDARDIZED EXPERIMENTS IN NUTRITION

Few aspects of experimental design in nutrition are currently standardized. Dietary interventions usually contain the variable of interest and yet there is no standardized requirement to state all of the ingredients. For example, fish oils have been routinely used as independent variables in nutrition studies, yet without a listing of the specific quantities of at least the major fatty acids of interest, ie, eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), the results are uninterpretable against other studies also investigating the effects of these fatty acids. Furthermore, there is no consensus for the composition of control diets.

If nutrition researchers formed consortiums or working groups to establish standards, multiple experiments would be far more comparable than they are today. Experimental outcomes for similar experiments often yield conflicting results that can be difficult to resolve. If experimental protocols, treatments, etc, were standardized, differing experimental outcomes would be more attributable to the variables of greater interest, such as dosage or time of intervention or the effect of the treatment on specific outcome measures. Standardization of nutrition experiments will enable individual researchers to contribute to larger research goals, obviating the need to form a large well-funded consortium for each new specific research agenda. Standards compliance will improve the accuracy of meta-analyses, fuel real progress in the ability to make dietary recommendations, and enable bioinformaticians and nutritionists to generate or test new hypotheses with data from prior experiments.

A recently published meta-analysis in this Journal provides a good framework for a discussion on the standardization of experimental design. This meta-analysis was conducted to determine the effect of n−3 long-chain polyunsaturated fatty acids (LC-PUFAs) on pregnancy outcomes and infant measurements at birth (17). Twenty-one studies were identified, 6 of which were used in the meta-analysis. Of these 6 studies, no 2 used the same treatments or controls or gestational duration of intervention. Even among the 4 studies that used an oil-based supplement treatment, the controls varied from corn oil to sunflower oil high in oleic acid to olive oil to an unidentified placebo (**Table 1**). These controls varied in content of poly- and monounsaturated fatty acids by as much as 80%.

TABLE 1
Example of standard treatments and controls

Original study	Original control	Original treatment	Standard control	Standard treatment
Helland et al (18)	Corn oil (10 mL/d)	Cod liver oil (10 mL/d); 1183 mg DHA, 803 mg EPA	DHA standard-matched control, 1000 mg/d	DHA standard, 1000 mg/d
Malcolm et al (19)	Sunflower oil (323 mg) with high concentrations of oleic acid	Fish oil; 200 mg DHA/d; < 40 mg EPA/d	DHA standard-matched control, 1000 mg/d	DHA standard, 1000 mg/d
Olsen et al (20)	Olive oil (4 g/d) or no oil supplement	Fish oil (4 g/d); 920 mg DHA/d + 1280 mg EPA/d	Fish oil–matched control, 4 g/d	Fish oil standard, 4 g/d
Sanjurjo et al (21)	Placebo	Fat (2 g/d); 200 mg DHA/d + 40 mg EPA/d	DHA standard-matched control, 1000 mg/d	DHA standard, 1000 mg/d
Smuts et al (22)	Ordinary eggs (18 mg DHA/egg); mean intake: 23 ± 9 mg DHA/d	High-DHA eggs; 135 mg DHA/egg; mean intake: 206 ± 112 mg DHA/d	Standard eggs; mean intake: ≈20 mg/d	Standard high-DHA eggs; mean intake: ≈200 mg/d
Smuts et al (23)	Ordinary eggs (33 mg DHA/egg); mean intake: 34 ± 16 mg DHA/d	High-DHA eggs; 133 mg DHA/egg; mean intake: 137 ± 65 mg DHA/d	Standard eggs; mean intake: ≈20 mg/d	Standard high-DHA eggs; mean intake: ≈200 mg/d

Many variations in study design involved in this example deserve further comment; however, for the purposes of illustrating the concept of what could be gained from standardization, the focus will be on the treatment and controls. Abbreviated descriptions of these aspects of study design from the original studies and how they might compare if standards were already in place are summarized in Table 1. The data in the table show that the process is imperfect from a meta-analysis standpoint because statisticians often pool studies that do not have the same original intention. Only half of the 6 original studies were designed with the intention to investigate DHA in an oil supplement form. One of the studies was devoted to marine oils and 2 of the studies were interested in DHA when delivered as part of egg intake. This is likely to be a widespread issue because nutrients may be investigated either in isolation or as part of a food matrix. Either type of investigation is legitimate, but both types of intakes should be standardized. Although some of the heterogeneity in the experimental design of similar experiments is due to genuine differences of intent with respect to the hypothesis being tested, much of it is still clearly a simple lack of consensus even when the intentions are identical.

The main conclusion of the particular meta-analysis reported (17) was that "n−3 LC-PUFA supplementation during pregnancy may enhance pregnancy duration and infant head circumference, but the mean effect size is small." Like many meta-analyses, the conclusion seems cautionary and somewhat inconclusive. Is there really an effect? If one calculates that every study costs several hundred thousand dollars, at least $4 million has already been spent on this question and yet the conclusion is incomplete.

If all or even a large percentage of the original 21 studies reviewed for the meta-analysis (17) followed the same experimental design, using the same standardized treatments and controls, then the statistical power of the combined studies would be enormously improved, and the conclusion as to whether or not an effect exists, and an estimate of the magnitude of the effect would likely be unequivocal. Furthermore, if the data from unpublished studies could also be used—those long forgotten in filing cabinets because the results were not desirable for publication—the last remaining bias (publication bias) would be removed. Regardless of the perceived desirability of the outcome, every standardized study would be a useful study, no funding dollars would be wasted, and progress in nutrition research would be exponentially improved.

DEVELOPMENT OF REPORTING STANDARDS FOR NUTRITION DATA

In order for nutrition data to be useful to other scientists for comparability or reproducibility, the data must be unambiguously described along with the experimental conditions, protocols, etc, that were used to generate the data. To some extent, these elements are described when the experimental results are published in a journal, but journals may have variable requirements and, because of length restrictions, some information may be abbreviated or omitted. Furthermore, to incorporate such information into a database, the data and the details of how the data were derived need to be described in a machine-readable format.

Fortunately, tremendous work toward the capturing of biological data types is already in progress. Beginning in 1999, an international organization of biologists, computer scientists, and analysts formed the Microarray Gene Expression Data (MGED) Society and created MIAME, a document that describes the minimum information that should be reported for each microarray experiment (10). Similar efforts are underway within the proteomics (24) and metabolomics (25) communities. The field of nutrition science can leverage these standards both by submitting their high-throughput data in these established formats and by using them as examples toward the development of reporting standards documents for data and experimental designs unique to nutrition.

The movement of nutrition data into databases, a major long-range goal listed in Figure 1, will require the development of reporting standards as a prerequisite. Such a specification or series of specifications would outline experimental results and associated data that are either unique to nutrition or not already described by a preexisting standard, such as MIAME. Whether or not measurements are from high-throughput experiments, uniform descriptions of the data and how they were produced are still needed to incorporate such data into a database. Nutrition research encompasses many different types of experimental designs, each with different types of data and associated annotations. Thus, separate specifications will likely be required for each type of experimental design in nutrition. Nonetheless, this

is precisely the stated goal for the future of nutrition as a field of integrated science (26).

Like the "-omics-driven" standards documents, the development of these reporting standards should be community-based to establish what information needs to be collected for each type of experiment. The collected information should be sufficient to enable replication of the experiment and to conduct comparisons with similar experiments. This step fully requires the leadership and participation of nutrition scientists. Additionally, the information should be structured in a way that will enable automated analysis and data mining (ie, with the use of nutrition ontologies), a task that requires the cooperation of both nutritionists and computer programmers.

DEVELOPMENT OF NUTRITION ONTOLOGIES

An ontology is a controlled vocabulary that also defines relations between the vocabulary terms. For example, in a hypothetical Ontology of Food Items, "apple" is a *type of* "fruit," whereas both "apple" and "fruit" are vocabulary terms in the ontology and *type of* is a relation description. A vocabulary is "controlled" if only a single term is used to express a given concept. For example, peptidylglycine monooxygenase, peptidyl α-aminating enzyme, and peptidylglycine-hydroxylase all confusingly refer to the same enzyme, but in the Enzyme Commission (EC) system, this enzyme has been assigned a single, controlled classification number: EC 1.14.17.3 (27). When scientists wish to unambiguously refer to an enzyme, the EC number is used.

Why are ontologies needed in nutrition? Conformance to a controlled vocabulary enables the automated analysis of experiments using bioinformatics. For example, if an experimental condition is described as "overnight fast," "12-h fast," or "fasted state" in different experiments, a researcher can manually decide that these are equivalent, but it is error-prone to do this with a computer. If a user queried the database for "fast," they might also get matches for "fast food."

Likewise, the formal description of relations enables computer-aided exploration of more abstract concepts. Even though a nutrition researcher doing an analysis knows that apples are types of fruit, computers do not know this a priori. If relations are defined, then automated analyses can be done based on those relations. For example, a nutrition researcher who is investigating the association of lung cancer with every type of food in an ontology-based food-frequency questionnaire (FFQ) could do an analysis with "apple" intake alone or with all food items that have been identified as a *type of* "fruit." Furthermore, if researchers would like to examine relations between different studies that used ontology-based FFQs, they would not have to manually or imperfectly resolve interstudy FFQ differences before conducting an analysis.

What types of ontologies are needed in nutrition research? Every piece of data or meta-data that is text-based needs to conform to a controlled vocabulary. Ideally, relations between the terms would also be defined. Preexisting standards such as the US Department of Agriculture's National Nutrient Database for Standard Reference (28) can be converted to ontologies. For example, using terms from this reference, "apples, raw, with skin" is a *type of* "fruit and fruit juice." Use of terms defined in this database would enable bioinformaticians to automatically link these foods to nutrient information. Multiple ontologies describing food items, experimental conditions, protocol descriptions, metrics, and so forth will be necessary to fully describe all aspects of experimental design in nutrition.

DEVELOPMENT OF NUTRITION DATABASES

What are the requirements for databases in nutrition? Unlike the sequence-centric databases that have been developed in molecular biology (29), in which queries are based on nucleic acid or amino acid sequences, nutrition databases would need to handle disparate types of data. Data structures need to be tailored for each data type, and the overall schema must be flexible and extendible, ie modular. The database structure must be able to handle detailed meta-data (ie, data that describe other data) using predefined controlled vocabularies or ontologies. The front-end interface must be user-friendly and, ideally, allow for functionally or physiologically based and sequence-centric queries. Finally, the database needs to be optimally integrated with other databases, because the overall knowledge repository related to each piece of data is dynamically changing. Ultimately, the design of databases by computer scientists will be largely dependent on completion of the reporting standards by nutrition scientists (Figure 1).

Reporting standards describe the content of the data and associated annotations. They do not specify the format in which the data are transferred, either between users and databases or between different databases. After nutrition researchers specify which data they want to submit to databases in the reporting standards, software developers then specify standard data-exchange formats that detail how the data will be transferred into and out of nutrition databases. So far, all of the data-exchange formats developed for high-throughput data have been based on an extensible markup language. For example, the data-exchange format developed for microarray data are MAGE-ML (Microarray Gene Expression Markup Language) (30). Once the specifications for data-exchange formats are complete, programmers can then build web interfaces and data submission tools to help nutrition researchers convert their data to the format understood by the database (**Figure 3**). From the perspective of nutrition researchers, the only visible component would be the questionnaire they answer about their data on the web.

FORMATION OF A NUTRITION STANDARDS BODY

Standards development for nutrition research would be greatly facilitated by the formation of a standards body or organization. A proposed agenda for such an organization is summarized as follows:

1) Develop standards for all aspects of experimental designs in nutrition.

2) Develop reporting standards so that data can be imported into databases.

3) Recruit computer programmers to develop data exchange formats and design databases.

4) Advocate adoption of new standards.

The development of microarray standards by the MGED Society is a good example of how this can be done from a logistical standpoint, as the organization began as a grass-roots movement without any dedicated funding. Meetings can be held in conjunction with major nutrition conferences, with individuals attending the relevant working group and each working group dedicated to a particular experimental design type (eg, epidemiology and

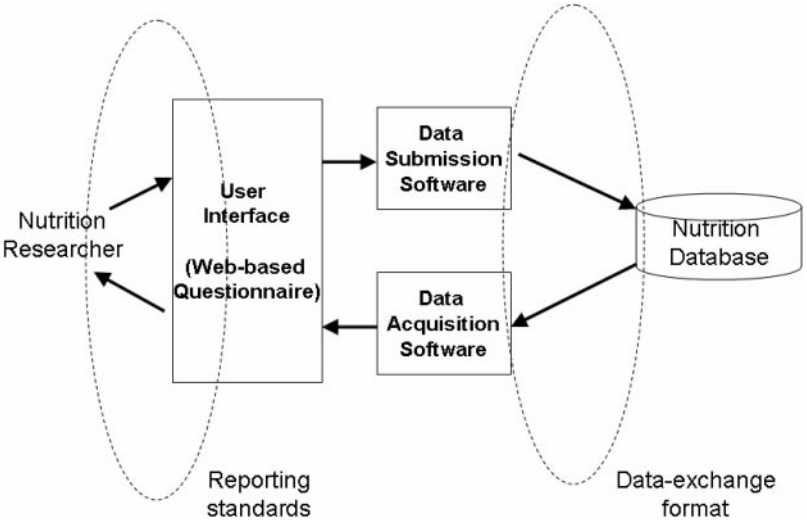

FIGURE 3. Specifications required to migrate nutrition data into databases. Reporting standards describe the data nutrition researchers submit to and retrieve from nutrition databases using Web-based questionnaires. Data-exchange formats, developed by computer scientists, describe the machine-readable format of these data. When these specifications are complete, computer programmers can create the software, depicted in the rectangles, that submits and retrieves the data from the database.

clinical trials). E-mail lists or Web-based tools for collaboration, such as wikis, can be established to facilitate communication and progress between meetings.

Although standards development and utilization may appear tedious for the individual researcher, just as a company would rather build a product however they want, such compliance is necessary for the future of the field of nutrition, just as it was necessary for the paradigm change of part interoperability in industry. Nutrition scientists who do not participate may find themselves left behind, whereas others who develop and utilize standards will increase the accessibility and impact of their data and, thus, their work will be well-cited and its impact proportionately greater.

EXAMPLE 1: DATABASES, ONTOLOGIES, AND AUTOMATION IN LARGE PROSPECTIVE TRIALS

The European Prospective In Cancer and Nutrition (EPIC) study is a large prospective study that, in 1993, began tracking half a million people in 10 countries at 23 different centers to find associations of dietary intake with cancer incidence (31). Questionnaire-style dietary intake measurements were collected from participants, and cancer incidents were tracked over time. As of 2006, only a small fraction of all of the possible associations between a dietary component and the incidence of a certain type of cancer has been published, with many studies investigating a single association. Why are whole manuscripts being devoted to a single statistical assessment?

When reviewing the study design and data collected in the EPIC trials (31), it appears that no 2 centers used the same dietary intake measurement. Had a standardized FFQ been applied at all study sites, it would have been technologically feasible to examine the association of every possible combination of food item or food group (groups of food items defined in a nutrition ontology) with every type of cancer via automated statistical analysis and present the findings in a single manuscript. This nonhypothesis-driven approach would have likely discovered unexpected and enlightening significant associations.

How can comprehensive, nonhypothesis-driven analyses of data from large prospective trials be conducted in the future? There are several key requirements for studies similar to the one described above: *1*) uniform food intake measurements across all study participants, *2*) compliance of food intake data to vocabulary words that have been defined in a nutrition ontology, *3*) storage of food intake and cancer incidence data in a centralized database, and *4*) automation of the statistical analysis to compute all possible statistically significant correlations. An additional benefit of such methodology is that the analysis can easily be recomputed when more cancer incidence data become available on these subjects.

Fortunately, there will be more opportunities to conduct data collection and analysis correctly in the future. Collins (32) has proposed a large US prospective study to discover gene-environment interactions in approximately half a million people who will be genotyped. This is a golden opportunity for nutritionists to uncover more gene-diet interactions, if the nutrition data are thoroughly and uniformly collected. What aspects of nutrition will be recorded and analyzed in this study? If the standardization and automation measures listed in the previous paragraph are implemented and the bioinformaticians responsible for the diet-related analyses are involved in database design from the beginning of the study, an unprecedented amount of additional information will be extracted from these samples. Additionally, the Collins study can provide substantially more analytic possibilities if appropriate biomarkers for nutrient status and phenotype are assessed in addition to survey-based food intake metrics.

EXAMPLE 2: RECYCLING NUTRITION DATA TO INVESTIGATE NEW HYPOTHESES

When nutrition data becomes more widely accessible, new hypotheses will be explored without conducting additional experiments. In this hypothetical example, a nutrition researcher hypothesizes that the effect of DHA intake on plasma fatty acid concentrations in pregnant women is significantly different from

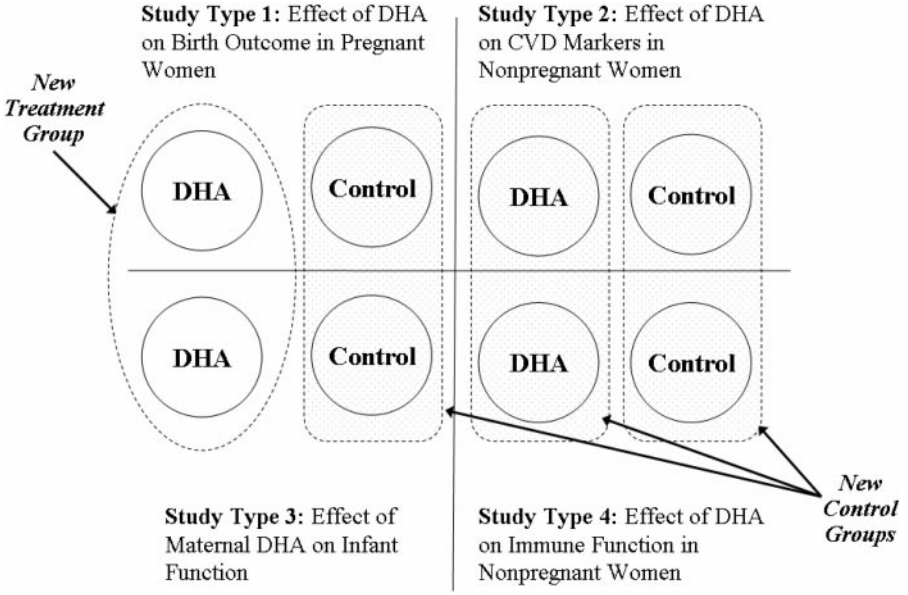

FIGURE 4. Investigating a new hypothesis with the use of data from previous studies. In this example, data from 4 different types of published experiments on docosahexaenoic acid (DHA) supplementation using pregnant and nonpregnant women were collected, and an intermediate data point that was commonly assessed in the experiments (eg, plasma fatty acid concentrations) was analyzed to explore pregnancy-related differences. The circles depict the original treatment and control groups within each of the 4 types of studies, whereas the dashed squares depict the new treatment and control groups created from the original data. These new treatment and control groups can now be compared with ANOVA analyses to test the new hypothesis with the existing data. CVD, cardiovascular disease.

the effect of DHA intake in nonpregnant women. To investigate this hypothesis, researchers collect various studies on DHA supplementation in pregnant and nonpregnant women, such as studies that explore the effect of DHA supplementation on pregnancy outcomes, on birth outcomes, on cardiovascular disease, on immune function, and so forth, in which measurements of plasma fatty acid concentrations were recorded. Assuming that the raw data from all of these previous studies are available and that the experiments conformed to standardized designs, researchers could then examine this new question by re-clustering all of the data, as shown in **Figure 4**, from the various studies into 4 separate groups: DHA-supplemented pregnant women, DHA-supplemented nonpregnant women, nonsupplemented pregnant women, and nonsupplemented nonpregnant women. ANOVA tests could then be applied to determine whether statistically significant intergroup differences exist. This type of computational hypothesis testing can greatly speed research progress. Even if this computational experiment is inconclusive, it can still provide information that can guide the study design of future experiments (eg, sample size requirements).

EXAMPLE 3: POOLING NUTRITION DATA FOR BIOMARKER DISCOVERY

When nutrition data based on standardized experimental design are accessible in public databases, bioinformaticians can pool data from multiple studies for various aims, such as hypothesis testing and biomarker discovery. Biomarkers are biological characteristics that distinguish specific biological states such as "healthy compared with metabolic syndrome" or "normal diet compared with calcium-supplemented diet." The use of biomarkers in nutrition studies is especially valuable as a companion to food intake measures because current techniques to measure food intake are based largely on subjective, biased reports provided by the participant. Independent, objective biochemical and physiologic markers for consumption of certain foods or nutrients provide independent validations of subject reports.

Biomarker discovery is not new to nutrition, but nutrition studies conducted to date appear to be limited to the identification of either singular biomarkers (eg, cholesterol) or multiple biomarkers from small numbers of biological samples. For example, researchers screened the plasma of healthy human subjects before and after supplementation with α-tocopherol and identified 12 peptides that differed by at least 2-fold (33). Another group of researchers studied the effect of vitamin C supplementation on plasma proteins in hemodialysis patients, and found 15 peptides that differed before and after supplementation (34). Unfortunately, neither of these 2 studies provides a usable biomarker profile because the differentiating patterns were not tested in the classification of independent samples, presumably because of the limited number of samples. However, the results do encouragingly suggest that vitamin status has effects discernable by analysis of the plasma proteome and that biomarkers could be identified if samples could be pooled from multiple standardized studies to increase sample size.

A hypothetical scenario illustrates how studies can be pooled to develop multiple biomarkers or bioprofiles. The subset of studies collected in the previous example (*see* Figure 3) and in other studies involving DHA supplementation in human subjects that include gene expression data could be collected to look for a transcriptomic bioprofile that distinguishes between DHA-supplemented and nonsupplemented subjects. A classification algorithm, reviewed by Kapetanovic et al (35), would then be applied to the data from a subset of the available samples, ie, a training set, to generate a biomarker profile or bioprofile that distinguishes the DHA-supplemented samples from the control

samples. Next, to determine the predictive capability of the identified biomarkers, the trained algorithm would be applied to an independent data set, ie, the remaining samples that were not used to train the algorithm. Because classification algorithms are not data type–specific, the incorporation of additional data sets (eg, proteomic and metabolomic) can improve the specificity of the resulting bioprofile.

Lessons from cancer research suggest that large numbers of human subjects are needed to identify biomarkers. Mukherjee et al (36) analyzed 8 cancer classifier studies of varying degree of classification difficulty to estimate sample size requirements for microarray analysis in human studies. As expected, an increase in sample size increased the accuracy of the classifier, but the number of human samples required to accurately predict treatment outcome was quite high, ie, 75–100 samples. For successful biomarker discovery in nutrition, the pooling together of larger numbers of human samples from multiple controlled studies will be a pragmatic necessity, and therefore, will be dependent on the standardization of experiments in nutrition.

CONCLUSIONS

An obvious prerequisite to the use of informatic tools on nutrition data are the accessibility of those data. As a first step, major journals in nutrition can immediately affect the field by requiring the online submission of data sets as a requirement for publication in their journal. Meanwhile, longer-range goals of standards development should be pursued by a nutrition standards body to enable the storage of nutrition data in databases and interstudy analyses. With nutrition data accessible in electronic formats, researchers will be able to generate or investigate hypotheses using sophisticated and powerful computational tools. Furthermore, if crucial aspects of the experimental designs have been standardized based on expert-defined criteria, the data will be readily comparable, improving the confidence in the findings and enabling previously impossible discoveries. Such revolutionary changes in the approach to hypothesis generation and testing have the potential to drastically accelerate progress in nutrition research.

The authors acknowledge the editorial assistance of CJ Dillard. All authors contributed to the preparation of this manuscript. The authors had no conflicts of interest.

REFERENCES

1. Lemay DG, Hwang DH. Genome-wide identification of peroxisome proliferator response elements using integrated computational genomics. J Lipid Res 2006;47:1583–7.
2. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 1999;15:563–77.
3. Zhou Q, Liu JS. Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 2004;20:909–16.
4. Hinrichs AS, Karolchik D, Baertsch R, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acid Res 2006;34:D590–8.
5. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000; 25:25–9.
6. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 2003;4:R7.
7. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acid Res 1996;24:238–41.
8. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acid Res 2003;31:3576–9.
9. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994;2:28–36.
10. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001;29:365–71.
11. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acid Res 2005;33: D562–6.
12. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2007;35:D21–5.
13. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acid Res 2007;35:D26–31.
14. The Universal Protein Resource (UniProt). Nucleic Acid Res 2007;35: D193–7.
15. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acid Res 2000;28:235–42.
16. Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council. Sharing publication-related data and materials: responsibilities of authorship in the life sciences. Washington, DC: The National Academies Press, 2003.
17. Szajewska H, Horvath A, Koletzko B. Effect of n−3 long-chain polyunsaturated fatty acid supplementation of women with low-risk pregnancies on pregnancy outcomes and growth measures at birth: a meta-analysis of randomized controlled trials. Am J Clin Nutr 2006;83:1337–44.
18. Helland IB, Saugstad OD, Smith L, et al. Similar effects on infants of n−3 and n−6 fatty acids supplementation to pregnant and lactating women. Pediatrics 2001;108:E82.
19. Malcolm CA, McCulloch DL, Montgomery C, Shepherd A, Weaver LT. Maternal docosahexaenoic acid supplementation during pregnancy and visual evoked potential development in term infants: a double blind, prospective, randomised trial. Arch Dis Child Fetal Neonatal Ed 2003; 88:F383–90.
20. Olsen SF, Sorensen JD, Secher NJ, et al. Randomised controlled trial of effect of fish-oil supplementation on pregnancy duration. Lancet 1992; 339:1003–7.
21. Sanjurjo P, Ruiz-Sanz JI, Jimeno P, et al. Supplementation with docosahexaenoic acid in the last trimester of pregnancy: maternal-fetal biochemical findings. J Perinat Med 2004;32:132–6.
22. Smuts CM, Borod E, Peeples JM, Carlson SE. High-DHA eggs: feasibility as a means to enhance circulating DHA in mother and infant. Lipids 2003;38:407–14.
23. Smuts CM, Huang M, Mundy D, Plasse T, Major S, Carlson SE. A randomized trial of docosahexaenoic acid supplementation during the third trimester of pregnancy. Obstet Gynecol 2003;101:469–79.
24. Orchard S, Hermjakob H, Taylor CF, et al. Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17–20th April 2005). Proteomics 2005; 5:3552–5.
25. Bino RJ, Hall RD, Fiehn O, et al. Potential of metabolomics as a functional genomics tool. Trends Plant Sci 2004;9:418–25.
26. Zeisel SH, Allen LH, Coburn SP, et al. Nutrition: a reservoir for integrative science. J Nutr 2001;131:1319–21.
27. Bairoch A. The ENZYME database in 2000. Nucleic Acid Res 2000; 28:304–5.
28. US Department of Agriculture, Agricultural Research Service. 2006. USDA National Nutrient Database for Standard Reference, release 19. 2006. Internet: http://www.ars.usda.gov/ba/bhnrc/ndl (accessed 13 February 2007).
29. Galperin MY. The molecular biology database collection 2006 update. Nucleic Acid Res 2006;34:D3–5.
30. Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002;3:RESEARCH0046.

31. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr 2002;5:1113–24.
32. Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429:475–7.
33. Aldred S, Sozzi T, Mudway I, et al. Alpha tocopherol supplementation elevates plasma apolipoprotein A1 isoforms in normal healthy subjects. Proteomics 2006;6:1695–703.
34. Weissinger EM, Nguyen-Khoa T, Fumeron C, et al. Effects of oral vitamin C supplementation in hemodialysis patients: a proteomic assessment. Proteomics 2006;6:993–1000.
35. Kapetanovic IM, Rosenfeld S, Izmirlian G. Overview of commonly used bioinformatics methods and their applications. Ann N Y Acad Sci 2004; 1020:10–21.
36. Mukherjee S, Tamayo P, Rogers S, et al. Estimating dataset size requirements for classifying DNA microarray data. J Comput Biol 2003;10:119–42.

The American Journal of Clinical Nutrition