Interoperating DNA Gene Sequences and Nutrition Provisions for Personalized Wellness

Jong P. Yoon[1] and Joyce Yoon[2]

[1] Mercy College, MATH/CIS Dept, Dobbs Ferry, New York
[2] Columbia University, Institute of Human Nutrition, College of Physicians and Surgeons New York, New York
jyoon@mercy.edu, jny2106@columbia.edu

Abstract. The last decade has seen a large explosion of health-related human centered computing research and practice focused on wellness such as provisioning good nutrition. Human genome projects become successfully recognized and DNA sequencing technologies become affordable. While Health Informatics may appear to be the obvious home for these activities, interoperability between DNA gene sequence analysis and nutrition support is less satisfactorily achieved. To promote humans health and wellness, this paper describes two-tied nutrition support system: rule-based and case-based reasoning about gene sequences and nutrients. In addition, this paper also protects human information from unauthorized accesses while preserving its privacy.

Keywords: Nutrition-net, Gene-net, Privacy Preservation, Rule-based Gene Sequence and Nutrient Reasoning, Case-based Reasoning.

1 Introduction

The last decade has seen a large explosion of health-related human centered computing research and practice focused on wellness such as good nutrition with the intention of helping people avoid needing medical care. Human genome projects become successfully recognized and DNA sequencing technologies become affordable [2,8,10]. While Health Informatics may appear to be the obvious home for these activities [3], it needs to provide the highest quality nutritional care for human wellness.

Malnutrition is not only nutritional imbalance or over nutrition, but also irrelevant or inaccurate nutrition. Nutrition usually refers to a deficiency of nutrients (under nutrition) relative to body requirements which contributes to an abnormality in body composition and/or its function [6,7]. This deficiency may arise from inadequate intake or absorption of good food, or exceeding intake of improper foods to a specific human body. It is likely that a food is good to a body while harm to another.

The resources listing accurate nutrition information is available nationwide and worldwide as well, with smart eating guidelines (e.g., http://www.nutrition.gov, http://www.nal.usda.gov/fnic/pubs/bibs/gen/eatsmart.pdf, or http://fnic.nal.usda.gov /resourcelists). It is important to eat healthy foods but more importantly to avoid the

J. Suzuki and T. Nakano (Eds.): BIONETICS 2010, LNICST 87, pp. 325–333, 2012.

foods which are adverse to the body state. Human body contains chemical compounds, such as water, carbohydrates (sugar, starch, and fiber), amino acids (in proteins), fatty acids (in lipids), and nucleic acids (DNA and RNA). These compounds are composed of elements such as carbon, hydrogen, oxygen, nitrogen, phosphorus, calcium, iron, zinc, magnesium, manganese, and so on. All of these chemical compounds occur in various forms and combinations (e.g., hormones, vitamins, phospholipids, hydroxyapatite) in the human body.

The contribution of this paper includes 1) personalized wellness promotion that can be achieved from interoperation between gene sequence analysis and nutrition development, and 2) a security technique that can protect human information from unauthorized accesses and providing services by preserving the privacy of human information. A broad architecture of the proposed technique is as shown in Figure 1. For wellness of human life, a list of foods is provisioned to a person based upon his or her DNA sequence. This provision is made to more general group of people (④ in Figure 1), and further specifically to a specific person (⑤ in figure). Since there exists the privacy issue in ⑤, a person can communicate securely.

The remainder of this paper is organized as followed: Section 2 describes about nutrition and nutrition-net. Section 3 describes about DNA and gene-net. Section 4 describes

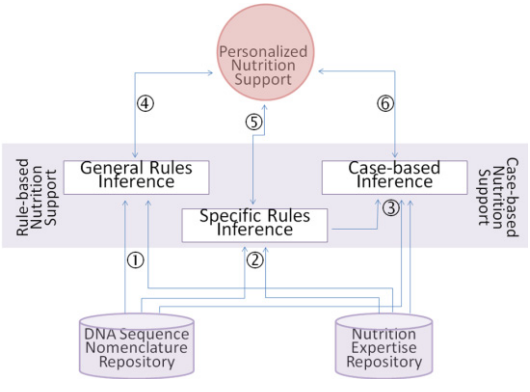

Fig. 1. General System Architecture

2 Nutrition

Nutrition is the provisioning process of the materials necessary foods to the organisms or cells of body [6,10]. Three are six major classes of nutrients: water, carbohydrates, protein, vitamin, fats and minerals. Foods are those edible substances that can provide human body with certain kinds of nutrients, like proteins, fat, fiber, carbohydrates, and different kinds of vitamins and minerals.

The stomach has limited space and should only be filled foods that are rich in essential nutrients, and are therefore beneficial for the improvement and development

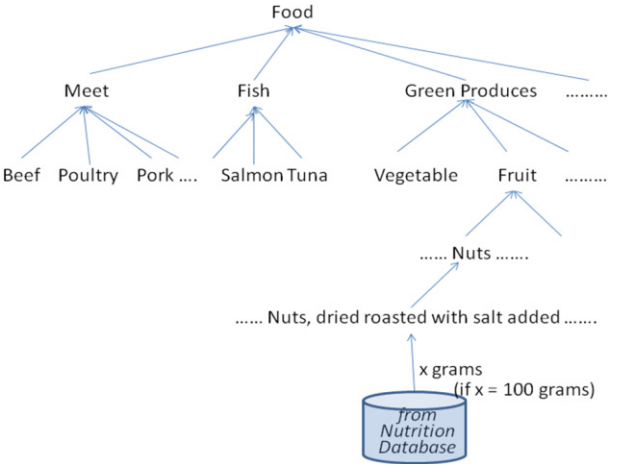

Fig. 2. Food Hierarchy of Nutrition-Net

of human brain and body. Incorporating additive nutrients into our food products is not new at all. Milk producers have been fortifying milk by adding extra vitamin D to it since the 1930s. Vitamin D helps prevent rickets in children, who are most vulnerable to this disease. For people who aren't able to consume lactose, adding vitamin D to their diets is a problem.

Also, the amount of nutrients that a human body can or need to consume is important. For example, more than 2000 milligrams of vitamin C per day in an adult human body can cause diarrhea and kidney stones. Stanols and plant sterols help reduce cholesterol inside one's body, and 2 grams of these substances per day is enough to keep cholesterol in control. Ingesting more than that amount can harm the body and does not provide any extra benefit. Iron is necessary for the growth of blood cells, but in some forms iron is not solvable is simple flushed out of the body.

Foods can be hierarchically organized [2]. One of the examples is shown in USDA National Nutrient Database (i.e., http://www.nal.usda.gov/fnic/foodcomp/search/). Foods are classified as being meet, fruits, vegetables, fishes, etc. Each can be also further specialized. For example, meats can be beef, pork, poultry, etc, each of which in turn can be specialized. Foods can be organized in a net, which we call "nutrition-net."

A nutrition-net is composed of a hierarchy of foods and the nutrition

Table 1. Nutrients Returned from the Nutrition Database

Nutrient	Value per 100 grams
Water	1.75g
Protein	17.30g
Lipid (fat)	51.45g
…	….
Carbohydrate	25.35g
Fiber, total dietary	9.0g
Sugars	4.65g
Calcium	70mg
…	…
Vitamin C	0.4mg
Vitamin B-6	0.296mg
Choline	54.3mg
….	….
Tryptophan	0.264g
Leucine	1.371g
….	….
Caffeine	0mg

database. As shown in Figure 2, the nutrition data may provide a list of the nutrients. In the figure, consider the food called “nuts, dried roasted with salt added” in the leaf node. It contains the nutrients as shown in Table 1 in the case of 100 grams of the food. Depending on the variable x, which indicates the weight of food, different values of nutrients will be provided. An example is listed as shown in Table 1.

3 DNA

Deoxyribonucleic Acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms with the exception of some viruses. DNA occurs as linear chromosomes in eukaryotes, and circular chromosomes in prokaryotes. The set of chromosomes in a cell makes up its genome. The human genome has approximately 3 billion base pairs of DNA arranged into 46 chromosomes. The information carried by DNA is held in the sequence of pieces of DNA called genes.

An alternative form of a gene, i.e., one member of pairs, located at a specific position on a specific chromosome is called an allele. Each gene can have different alleles. Sometimes, different DNA sequences or alleles can result in different traits, for example skin color. It is likely that different DNA sequences or alleles may have the same result in the expression of a gene. At each locus among various individuals, multiple alleles exist. Allelic variation at a locus is measurable as the number of alleles present or the proportion of heterozygotes in the population.

Polymerase Changing Reaction (PCR) is a technique in molecular biology to amplify a single or few copies of the piece of DNA across several orders of magnitude, generating thousands to millions of copies of a particular DNA sequence. The method relies on thermal cycling, consisting of cycles of repeated heating and cooling of the reaction for DNA melting and enzymatic replication of the DNA. Primers (short DNA fragments) containing sequences complementary to the target

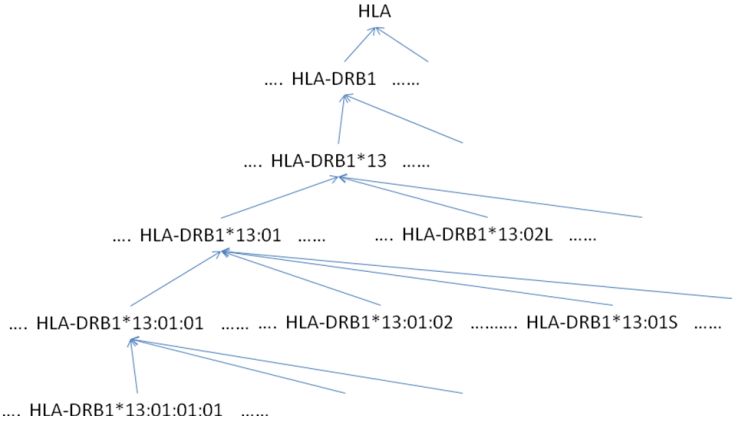

Fig. 3. An Example of Gene-Net

region along with DNA polymerase are key components to enable selective and repeated amplification. As PCR progresses, the DNA generated is itself used as a template for replication, setting in motion a chain reaction in which the DNA template is exponentially amplified.

Virtual Hybridization is a technique that rearranges chromosomes and enables to infor phylogenetic relationship between species using "gene" order. Human Leukocyte Antegen (HLA) is the name of the major histocompatibility complex in humans. The locus of the HLA contains a large number of genes related to immune system function in humans. This group of genes resides on chromosome 6, and encode cell-surface antigen-presenting proteins and many other genes.

Each HLA allele name has a unique number corresponding to up to four sets of digits separated by colons. The length of the allele designation is dependent on the sequence of the allele and that of its nearest relative. All alleles receive at least a four digit name, which corresponds to the first two sets of digits, longer names are only assigned when necessary. Modern HLA alleles are typically noted with a variety of levels of detail. Most designations begin with HLA- and the locus name, then * and some number of digits specifying the allele. The first two digits specify a group of alleles. Older typing methodologies often could not completely distinguish alleles and so stopped at this level. The third through fourth digits specify a synonymous allele. Digits five through six denote any synonymous mutations within the coding frame of the gene. The seventh and eighth digits distinguish mutations outside the coding region. Letters such L, N, Q, or S may follow an allele's designation to specify an expression level or other non-genomic data known about it. Thus, a completely described allele may be up to 9 digits long, not including the HLA- prefix and locus notation.

For example, the nomenclature HLA-DRB1*13 specifies a group of allels which encode the DR13 antigen or sequence homology to other DRB1*13 alleles, while HLA-DRB1*13:01 specifies a specific HLA allele. HLA-DRB1*13:01:02 specifies an allele that differs by a synonymous mutation from DRB1*13:01:01, while HLA-DRB1*13:01:01:02 specifies an allele which contains a mutation outside the coding region from DRB1*13:01:01:01. As exemplified, there are relationships amongst DNA gene or allele sequences. Such relationships can be depicted in a graph, called gene-net. An example of gene-net is shown in Figure 3.

4 Wellness Inference

As discussed in the previous sections, we know that foods can be organized in the nutrition-net, while DNA gene sequences are in a gene-net. In Figure 1, there are two types of logical expression: general expression and specific expression. A general expression is defined over general terms, while a specific expression over specific terms. For example, *Gene*("freshmen in CS dept", "UCC"), meaning that freshmen in the class CISC131 has a gene sequence codon UCC, is a general expression [9], and *Gene*("Adam", "UCC") is a specific expression. In this section, we do not distinguish them. However, Section 5 considers the situation where specific rules need to inference, and therefore, the privacy of a specific information needs to be preserved.

This section describes how to suggest a food to a person according to his or her body state. There are basically two cases of nutrition suggestion: well-known rule-based suggestion and case-based reasoning suggestion [8].

4.1 Rule-Based Suggestion for Well-Known Cases

Consider the two rules; 1) a certain gene sequence causes a symptom, and 2) food is good for those who have specific DNA gene sequences. Such rules can be written in logic as follows:

$$\begin{array}{l} \text{Gene } (s, g) \rightarrow \text{Symptom } (s, x) \\ \text{Food } (s, f, a) \rightarrow \text{Gene } (s, g). \end{array} \quad (\mathbf{1})$$

These two rules imply that if a person s has the gene sequence g, then that person has the symptom of (disease) x. Also, the gene sequence g is caused if a amount of food f is consumed.

From the rules in (1), the following rule is derived.

$$\frac{\text{Gene } (s, g) \rightarrow \text{Symptom } (s, x), \text{ and Food } (s, f, a) \rightarrow \text{Gene } (s, g)}{\text{Gene } (s, g) \rightarrow \neg \text{ Food } (s, f, a)} \quad (2)$$

meaning that if the two rules, Gene (s, g) → Symptom (x) and Food (s, f, a) → Gene (s, g), are given, the food f has mal-nutrient.

For example, it is known that the A1 allele associated with DRD2 gene binding in the striatum and compromised striatal dopamine signaling reduces obese individuals. In other words,

$$\begin{array}{l} \text{Gene } (s, \texttt{"DRD2"}) \rightarrow \text{Symptom } (s, \texttt{"obesity"}), \text{ and} \\ \text{Food } (s, \texttt{"Sugar"}, a) \rightarrow \text{Gene } (s, \texttt{"DRD2"}). \end{array}$$

With these rules, it turns out that Sugar is not recommended to those who have gene "DRD2". That is, Gene (s, "DRD2") → Food $(s,$ "Sugar", $a)$.

We further consider both nutrition-net and gene-net. Since hierarchical, they can be also rewritten in logic as follows:

$$\frac{\text{Gene } (s, g) \rightarrow \text{Food } (s, f_i, a), \text{ and } f_i \rightarrow f_j}{\text{Gene } (s, g) \rightarrow \text{Food } (s, f_j, a).} \quad (3)$$

and

$$\frac{\text{Gene } (s, g_i) \rightarrow \text{Food } (s, f_i, a), \text{ and } g_i \rightarrow g_j}{\text{Gene } (s, g_j) \rightarrow \text{Food } (s, f_j, a).} \quad (4)$$

Equation (3) means that if rule Gene (s, g) → Food (s, f_i, a) exists and food f_j is super entity to another food f_i in a nutrition-net, then a new rule is generated such that Gene (s, g) → Food (s, f_j, a). Similarly, Equation (3) states that if rule Gene (s, g_i) →

Food (s, f_i, a) exists and gene g_j is a super entity to another gene g_i in a gene-net, then a new rule Gene (s, g_j) → Food (s, f_i, a) is generated.

4.2 Cased-Based Reasoning

We know that there are many undiscovered rules, which however will be yet another useful food suggestion to human life. For this we employ the technique of case-based reasoning [8]. The case-based reasoning is the process of solving new problems on the solution of similar past problems. The idea comes from the fact that an auto mechanic who fixes an engine by recalling another car that exhibited similar symptoms. For wellness of human life (in Figure 1-⑥), this case-based reasoning may be more frequently used in part because human DNA gene sequences are increasing and its new nomenclature is designed.

An example of case-based reasoning is depicted in Figure 4. We know how to suggest the healthy food N_i if gene sequence Ω_i exists. For the wellness for a person who has a gene sequence Ω_j, if there is no foods that are found healthy for the person, but we do know gene sequence Ω_i is similar to Ω_j as shown in the second diagram of Figure 4(2), then we may be able to suggest the food N_j. Of course, we need to have a transformation rule α_{ij} for gene sequence, and also β_{ij} for food transformation. Finding α_{ij} and β_{ij} is beyond the scope of this paper, but will be discussed in another paper.

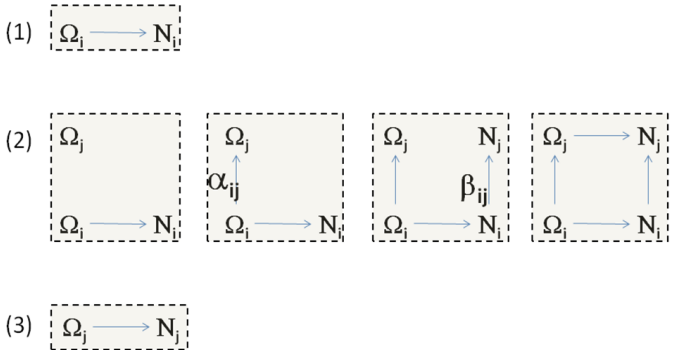

Fig. 4. Discovering in Case-based Reasoning

5 Privacy Issues

This section describes a technique that can preserve the privacy for particular personal information in specific rules inference. In Figure 1-⑤, since rules are defined over specific information, e.g., personal information, and since generalization approach [9] is not satisfied, user accesses need to be controlled properly. The access control techniques have been developed successfully [1]. However, they are not efficient in the domain interoperating two or more repositories. The following employs PKI [4].

To request the server to reason about specific DNA gene sequence with specific personal information in specific rules, participants need asymmetric keys. The following can generate private and public keys for nutritionist Joyce who helps patient John.

```
 keytool –genkey –alias CA1 –keystore herOnly.ks
–storepass ColumbiaUniversity –keypass nutrients.org
 –dname “CN=Joyce, OU=Nutrition, O=Columbia, L=New York,
ST=NY, C=US”;
```
(5)

After the certificate authority `CA1` signs on the request jar file `B.jar` containing DNA gene sequences for John, and to generate the signed jar file `Bs.jar`. The access policy is defined to grant the write permission to the jar code `Bs.jar` to a gene sequence database, `gen-net` to access John's gene sequence.

In this environment, the following java execution securely

```
java    –Djava.security.manager   –Djava.security.policy
= our.policy –cp Bs.jar John gen-net
```
(6)

In this way, the gene sequences for John and the nutrition data matched with his gene can be inter-shared securely and privately.

6 Summary

In this paper, we introduced two biological structures: DNA gene-net and nutrition-net. For interoperability of DNA gene-net and nutrition-net, two-tiered nutrition assistant system is proposed. The nutrition assistant system uses rule-based and case-based reasoning about to promote humans wellness and health. This interoperability are performed securely and privately.

References

1. Ben Ghorbel-Talbi, M., Cuppens, F., Cuppens-Boulahia, N., Bouhoula, A.: Managing Delegation in Access Control Models. In: IEEE ADCOM (2007)
2. Ghoting, A., Makarychev, K.: Indexing Genomic Sequences on the IBM Gene. In: Proc. of the Conf. on High Performance Computing Networking, Storage and Analysis (2009)
3. Haslhofer, B., Klas, W.: A survey of techniques for achieving metadata interoperability. ACM Computing Surveys 42 (2010)
4. Huang, J., Nicol, D.: A calculus of trust and its application to PKI and identity management. In: ACM IDtrust 2009 (2009)
5. Igel, C., Glasmachers, T., Mersch, B., Pfeifer, N., Meinicke, P.: Gradient-based Optimization of Kernel-Target Alignment for Sequence Kernels Applied to Bacterial Gene Start Detection, pp. 216–226 (2007)
6. Kelley, P., Bresee, J., Cranor, L., Reeder, R.: A “Nutrition label” for privacy. In: Proc. of the 5th Symposium on Usable Privacy and Security (2009)
7. National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov`

8. Neves, M., Chagoyen, M., Carazo, J., Pascual-Montano, A.: CBR-Tagger: a Case-based Reasoning Approach to the Gene/Protein Mention Problem. In: Proc. of the Workshop on Current Trends in Biomedical Natural Language Processing (2008)
9. Rowell, L.: Writing the rules of digital privacy. ACM networker 13(2) (2009)
10. Zhao, W., Serpedin, E., Dougherty, E.: Inferring Connectivity of Genetic Regulatory Networks Using Information-Theoretic Criteria. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 262–274 (2008)