Journal of
Data Mining in Genomics & Proteomics

Lisitsa et al., J Data Mining Genomics Proteomics 2015, 6:2
DOI: 10.4172/2153-0602.1000173

Review Article Open Access

# Is it Time for Cognitive Bioinformatics?

**Andrey Lisitsa[1,2*], Elizabeth Stewart[2,3], Eugene Kolker[2-5]**

[1]The Russian Human Proteome Organization (RHUPO), Institute of Biomedical Chemistry, Moscow, Russian Federation
[2]Data Enabled Life Sciences Alliance (DELSA Global), Moscow, Russian Federation
[3]Bioinformatics and High-Throughput Data Analysis Laboratory, Seattle Children's Research Institute, Seattle, WA, USA
[4]Predictive Analytics, Seattle Children's Hospital, Seattle, WA, USA
[5]Departments of Biomedical Informatics & Medical Education and Pediatrics, University of Washington, Seattle, WA, USA

**Abstract**

The concept of cognitive bioinformatics has been proposed for structuring of knowledge in the field of molecular biology. While cognitive science is considered as "thinking about the process of thinking", cognitive bioinformatics strives to capture the process of thought and analysis as applied to the challenging intersection of diverse fields such as biology, informatics, and computer science collectively known as bioinformatics. Ten years ago cognitive bioinformatics was introduced as a model of the analysis performed by scientists working with molecular biology and biomedical web resources. At present, the concept of cognitive bioinformatics can be examined in the context of the opportunities represented by the information "data deluge" of life sciences technologies. The unbalanced nature of accumulating information along with some challenges poses currently intractable problems for researchers. The solutions to these problems at the micro-and macro-levels are considered with regards to the role of cognitive approaches in the field of bioinformatics.

**Keywords:** Digital medicine; Exposome; Omics; Biomarkers

## Deluge of Life Sciences Big Data

Modern molecular biology studies the function and interactions of biological molecules, processes and systems. Molecular biology primarily focuses on DNA, RNA, proteins and metabolites utilizing four corresponding „omics" technologies: genomics, transcriptomics, proteomics, and metabolomics. Advances in sequencing technologies enable researchers not only to decipher the human genome, but also to uncover features of epigenetic regulation. The level of gene expression is the object of transcriptomics; proteomics provides information about the diversity of the protein molecules encoded by the genome; metabolomics focuses on studying metabolites.

The technical capabilities of modern „omics" significantly exceed the capabilities of researchers to meaningfully process incoming experimental information. A typical microarray study can generate the data on 500,000 single-nucleotide substitutions of DNA that differentiate the genome of one person from another [1]. Transcriptome analysis allows measuring the activity levels of each of 20,000 human protein-coding genes, with the signal of each gene detected by two or three different sequences [2]. An in-depth study of the proteome can quantify up to 10,000 protein products of the genome [3] and elucidate structural modifications of at least one third of them [4].

Current international projects are designed not only to read the DNA sequence, but also to decode the meaningful message from this molecule. The question of how to re-interpret the genome in the context of harmful mutations and diseases is of great interest to biomedical researchers [5]. Such projects that are the sequels of the Human Genome Project include: HapMap, EnCode, 1000 Genomes, Cancer Genome, Human Proteome and a number of other large-scale initiatives. The projects like the NIH Brain Initiative, the Human Microbiome Project are generating the data that will build a framework for understanding the human condition. These projects combine the technical and intellectual capacities of numerous countries into international efforts.

There is no lack of baseline data on the state of molecular systems within a human body. On the contrary, the exponentially growing stream of the information produced by "omics" is a challenge to the technology of data storage and processing [6]. The unbalanced nature of accumulating information redoubles the challenges of the 5Vs of Big Data: veracity/reproducibility [7], variety, value and, to a lesser extent, velocity and volume [8].

The challenge of interpreting the genome is currently being tackled in many labs by using brute-force algorithms, making exhaustive searches for genetic variants, the laborious matching of scattered IDs and hand-curating datasets. However, the data assembled from all sources are left to languish in repositories "till better times" until the development of bioinformatics grows to the capability of gaining knowledge from the data sets accumulated.

## Challenges and Opportunities

Is there a background in modern bioinformatics to beat the challenge of current data deluge? In 2004, Kuchar et al. [9] introduced cognitive bioinformatics as an instrument of scenario analysis performed by scientists working with molecular biology and biomedical web resources. The term "Big Data" is mainly used to describe a massive volume of both structured and unstructured data that is too large or complex to process it using traditional database and software techniques. Bioinformatics accumulates the data primarily borrowed from the statistics, machine learning and pattern recognition methods (except for the specific tasks of molecular modeling and

***Corresponding author:** Lisitsa Andrey, The Russian Human Proteome Organization (RHUPO), Institute of Biomedical Chemistry, Moscow, Russian Federation, Tel: +7(499)246-69-80; Fax: +7(499)245-08-57; E-mail: lisitsa063@gmail.com

**Received** June 05, 2015; **Accepted** July 07, 2015; **Published** July 14, 2015

**Citation:** Lisitsa A, Stewart E, Kolker E (2015) Is it Time for Cognitive Bioinformatics? J Data Mining Genomics Proteomics 6: 173. doi:10.4172/2153-0602.1000173

comparative genomics). Thus, the algorithms of digital data processing cannot be transferred to biological and biomedical outcomes without significant expert adaptation. It is not surprising, for example, that the modern practice of clinical diagnostics has hardly been updated with new biomarkers despite the petabytes of the data generated. Potential solutions can lie in an analysis of network interactions from the micro-level (biological molecules, e.g., [10]) to the macro-level (exposome, [11]).

## Small World of Genes and Proteins

Socially relevant diseases affecting almost all people are multigenic. Thus, association studies are to be conducted to establish the link between multiple variations and the risk of disease development (GWAS - genome wide association studies). Sometimes, it is relatively easy to determine variation/disease relationships in controlled environments. However, to establish a relationship between a set of variations and a disease is rather difficult [12] due to the genome variability interfering with environmental factors.

It is possible to reduce the level of genomic variability at the level of proteins. The proteins, as the final products of the genome, have more direct influence on the condition of the body and can serve as reliable diagnostic markers. This issue has been thoroughly studied in the field of proteomics for the last 15 years. However, clinically useful biomarkers that can impact human health are lacking despite the reported hundreds of potentially relevant diagnostic molecules [13].

A single biomarker can be fallible because proteins interact with each other. They also interact with genes forming a genomic regulatory network. Networks of molecular interactions are studied in the framework of interactomics. This term labels the biology to be investigated but does not describe an appropriate approach. Graph theory (Petri nets), Bayesian models, support vectors, and random forest are some of the approaches that have been used to describe the networks of molecular interactions [14]. At present, it is only known that an organizational framework of hubs structured around the interacting partners can be detected only in the composition of networks with more than 200,000 binary interactions [15]. But such scale of experimental measurements is not affordable.

Disease development is determined by multiple conditions and in no way depends solely on personal genome [16]. The factor of genetic predetermination of a disease is relatively small. Pathologies in a genetically viable organism are formed by the components of environment collectively called the exposome [11]. It is a combination of environmental factors that an organism has encountered along with some susceptibilities inherent in the organism, which in combination leads to a disorder.

Data deluge suggests that solutions should encourage and harness the future convergent synthesis of exterior and interior life factors. The prerequisites for such a generalization could appear from the study of connectome [17] - the model of the brain as an organ available for the study at the social, cellular and molecular levels. It is impossible to prove the unity of these structures – the systems and a cell as a "society" of molecules. At the present level of knowledge, current mathematical tools do not adequately reflect the fundamental properties of the social, neural and biological networks, especially their time-dependent multi-layering. The attempts to construct planar, tree-like architectures of biological systems have been exhausted. Thus, there is a need for fundamentally different, probably non-digital, information processing tools to succeed in multi-dimensional descriptions.

Although it is possible for a clinician to get an enormous amount of the data on internal molecular systems of the body, he/she usually lacks the detailed information with respect to the exposome. Generally, the exposome is captured as a personal health record based on often-faulty memories of patients. Recent developments in modern telecommunication sphere have resulted in an opportunity of collecting the exposome data. First of all, it can be periodically received from our communicators such as geolocations [18]. With respect to the health characteristics of an individual, the sequence of geolocation is no less informative than the DNA sequence in the genome [19]. The information kept in electronic financial systems (for example, the information on purchases) along with Google searches and Facebook activity can be used to accurately build the behavioral profile of an individual. Some wrist trackers can measure physical activity, food/drug intake and other health parameters and generate such profiles as well. To avoid faulty memories these profiles can then be shared with a physician for a more accurate accounting of the patient's exposome. In addition, these profiles can be shared with others or connected to a community of like-minded (www.myfitnesspal.com [20], www.mynetdiary.com [21]). Such detailed profiles can also be valuable for evaluating drug side effects as well [19].

Geo-data are becoming a source of the prediction of behavioral profiles leading to the development of socially relevant diseases: diabetes, pathologies of the cardiovascular system and depression [22]. Further development of these concepts depends on creating a mathematical approach and the appropriate tools to describe the multi-layered cognitive machines. Today, there are no published computational principles that can surmount the barrier between the statics of traditional neural network and dynamics of real biological system. A strong attempt was made by IBM which provided its Watson artificial intelligence for medical application [23,24].

The methods for quantifying the exposome in terms of the environmental factors are generally indirect. For example, if the taxi order is accepted with a delay less than 5 minutes, it could indicate an aggressive industrial environment. The less the delay is, the more the chance is that a person's current location is within the generally unsuitable environment with hard traffic, noise and air pollution. Hypothetically, it is also possible to measure even mental environment by analyzing the number of calls per hour/per day, voice timber in a conversation and the diversity of vocabulary, as has been recently implemented in the service "Okay, Google".

Of course, the usefulness of exposome data cannot be disclosed until the data will be collected and provided for an analysis. The good news is that there is no need to invest scientific money into the collection of exposome data. The process of collection is fueled by business and goes in a fully automatic mode. The challenge is to make the data collected available to science and to develop the artificial intelligence to tackle this data.

Multi-layer cognitive systems should be based on a detailed description of the surrounding world. Artificial intelligence will perceive the physiological and emotional status of a person only if a critical mass of people consciously and without any pressure shares information about their location, the genome and diseases. And many other things that today we prefer to hide. Modern society is not fully prepared to accept such relationships due to ethical requirements regarding the protection of personal data. In medicine, in addition to this issue, there is a considerable need for a physician who can turn biomedical information into life-changing decisions.

## Private Health Care

A physician never makes a diagnosis solely based on the results of in-vitro tests. Even when there is a patient with an off-the-scale cholesterol level, the physician will never diagnose without an additional research, including examination data and questioning the patient. A correct diagnosis always depends and will depend upon the professionalism of the physician - perhaps the only parameter that cannot be formalized within the standards of medical care.

The development of "omics" as the technologies providing redundant information on the subject is essentially antagonistic to the concept of medicine. Indeed, if in the future (not that far, by the way!) bionic implants transmit gigabytes of the data on the state of a single molecule, a physician is unlikely to be able to read, understand and use these data without pre-buffering and analytical processing. But, such analysis will be completely useless until it is possible to compare the data between people, which will provide a sufficient pool of training images and an assessment of a "human norm".

It should be assumed that future systems with a cognitive component will be in demand as an interface between a physician and a patient data deluge. At present, the abundance of the data on a patient (for example, her/his genome) is more likely to overwhelm a physician than help in decision making. Currently, modern bioinformatics is aimed at decreasing the complexity of "omics" data through data reduction. Yet during this reduction, a valuable, individual, personalized picture disappears. A good contact between a physician and a patient in the array of "omics" data will be maintained only in case of the virtualization of both the physician and the patient. After virtualization the personal experience of decision-making of the physician can be copied repeatedly, resulting in the in silico "proliferation" of physicians. In the long term, the introduction of cognitive technologies in medicine should provide every citizen with a readily available ultra-professional virtual physician operating an infinite memory of "clinical records" with by hundreds of thousands of molecular parameters. The physicians as persons will become supervisors of a network of medical community and the most complex cases that require human intervention will rise to their level.

## Conclusions

Immense datasets are being generated by modern "omics" technologies to adequately represent and understand living organisms. Yet these datasets overwhelm the capabilities of any scientist, lab or research center to transform the data to actionable knowledge. Ironically and perhaps fortuitously, these datasets also contribute to our understanding of the brain and its cognitive processes and thus can be used to enhance our capabilities. Perhaps cognitive bioinformatics can formalize the logic of human thinking and apply it to organize the existing arrays of experimental data in the form of networks with dynamic behavior. The principles of networking between the neurons of the brain are likely to become scalable both from the macro-level to the level of social networks and from the micro-level to the level of intermolecular interactions in a cell. By applying one of the wonders of nature — the thought process — we can build this framework and analyze the approaches necessary to understand its secrets.

### References

1. The International HapMap Consortium (2004) Integrating ethics and science in the International HapMap Project. Nature Genetics 5: 467-475.
2. Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, et al. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol 7: 548.
3. Mann M, Kulak NA, Nagaraj N, Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. Molecular cell 49: 583-590.
4. Kelleher NL (2012) A cell-based approach to the human proteome project. J Am Soc Mass Spectrom 23: 1617-1624.
5. Tejedor JR, Valcárcel J (2010) Gene regulation: Breaking the second genetic code. Nature 465: 45-46.
6. Pennisi E (2011) Human genome 10th anniversary-Will computers crash genomics? Science 331: 666-668.
7. Collins FS, Tabak LA (2014) NIH plans to enhance reproducibility. Nature 505: 612-613.
8. Roger H, Haynes W, Stanberry L, Stewart E, Yandl G, et al. (2013) Unraveling the Complexities of Life Sciences Data. Big Data 1: 42-50.
9. Kuchar OA, Reyes-Spindola JF, Benaroch M (2004) Cognitive bioinformatics: computational cognitive model for dynamic problem solving. In Cognitive Informatics. Proceedings of the Third IEEE International Conference on, Canada.
10. Gstaiger M, Aebersold R (2013) Genotype–phenotype relationships in light of a modular protein interaction landscape. Mol Biosyst 9: 1064-1067.
11. Brunekreef B (2013) Exposure science, the exposome, and public health. Environ Mol Mutagen 54: 596-598.
12. Kaiser J (2012) Genetic influences on disease remain hidden. Science 338: 1016-1017.
13. Veenstra T (2011) Where are all the biomarkers? Expert Rev. Proteomics 8: 681–683.
14. Shoemaker BA, Panchenko AR (2007) Deciphering protein–protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43.
15. Cho YR, Zhang A (2010) Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. BMC Bioinformatics 11: S3.
16. Bartol J (2013) Re-examining the Gene in Personalized Genomics. Science and Education 22: 2529-2546..
17. Toga AW, Clark KA, Thompson PM, Shattuc DW, Van Horn JD (2012) Mapping the human connectome. Neurosurgery 71: 1-5.
18. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the Crowd: The privacy bounds of human mobility. Sci Rep 3: 1376.
19. White RW, Tatonetti NP, Shah NH, Altman RB, Horvitz E (2013) Web-scale pharmacovigilance: Listening to signals from the crowd. J Am Med Inform Assoc 20: 404-408.
20. https://www.myfitnesspal.com/
21. http://www.mynetdiary.com/
22. Pentland A, Lazer D, Brewer D, Heibeck T (2009) Using reality mining to improve public health and medicine. Stud Health Technol Inform 149: 93-102.
23. IBM and Apple Expand Partnership to Help Transform Medical Research. IBM news releases, 2015.
24. IBM Watson at Work: Transforming Healthcare - Boston Children's Hospital, Innovation Summit, 2014.