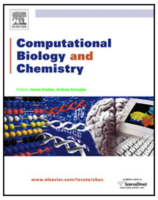
Computational Biology and Chemistry 47 (2013) 37–47

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Review article

Computational intelligence techniques in bioinformatics

Aboul Ella Hassanien[a,d,*], Eiman Tamah Al-Shammari[b], Neveen I. Ghali[c,d]

[a] *Faculty of Computers and Information, Cairo University, 5 Ahmed Zewal Street, Orman, Giza, Egypt*
[b] *Kuwait University, Faculty of Computing Science and Engineering, Kuwait*
[c] *Faculty of Science, Al-azhar University, Cairo, Egypt*
[d] *Scientific Research Group in Egypt (SRGE), Egypt*[1]

ARTICLE INFO

Article history:
Received 7 February 2013
Received in revised form 6 April 2013
Accepted 24 April 2013

ABSTRACT

Computational intelligence (CI) is a well-established paradigm with current systems having many of the characteristics of biological computers and capable of performing a variety of tasks that are difficult to do using conventional techniques. It is a methodology involving adaptive mechanisms and/or an ability to learn that facilitate intelligent behavior in complex and changing environments, such that the system is perceived to possess one or more attributes of reason, such as generalization, discovery, association and abstraction. The objective of this article is to present to the CI and bioinformatics research communities some of the state-of-the-art in CI applications to bioinformatics and motivate research in new trend-setting directions. In this article, we present an overview of the CI techniques in bioinformatics. We will show how CI techniques including neural networks, restricted Boltzmann machine, deep belief network, fuzzy logic, rough sets, evolutionary algorithms (EA), genetic algorithms (GA), swarm intelligence, artificial immune systems and support vector machines, could be successfully employed to tackle various problems such as gene expression clustering and classification, protein sequence classification, gene selection, DNA fragment assembly, multiple sequence alignment, and protein function prediction and its structure. We discuss some representative methods to provide inspiring examples to illustrate how CI can be utilized to address these problems and how bioinformatics data can be characterized by CI. Challenges to be addressed and future directions of research are also presented and an extensive bibliography is included.

Contents

---

* Corresponding author at: Faculty of Computers and Information, Cairo University, 5 Ahmed Zewal Street, Orman, Giza, Egypt. Tel.: +20 1114567956.
E-mail addresses: aboitcairo@fci-cu.edu.eg, aboitcairo@gmail.com (A.E. Hassanien), eiman.tamah@gmail.com (E.T. Al-Shammari), dr.eiman@ku.edu.kw (N.I. Ghali).

[1] http://www.egyptscience.net.

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.04.007

1. Introduction

During the past few decades we have experienced a massive growth in biological information gathered by the related scientific communities. A deluge of such information coming in the form of genomes, protein sequences, gene expression data and so on have led to the need for effective and efficient computational tools to store, analyze and interpret these data. Bioinformatics (Cios et al., 2005; Kelemen et al., 2008; Gusfield, 2004; Valentini et al., 2009; Smolinski et al., 2008a,b) involves the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems, often on the molecular level. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expressions and protein–protein interactions, and the modeling of evolution (Anon., 2013; Smolinski et al., 2008a,b). Hence, in other words, bioinformatics can be described as the application of computational methods to make biological discoveries (Baldi and Brunak, 1998; Mitra and Hayashi, 2006).

The ultimate attempt of the field is to develop new insights into the science of life as well as creating a global perspective, from which the unifying principles of biology can be derived (Altman et al., 2001). Cohen (2004) explained the needs of biologists to utilize and help interpret the vast amounts of data that are constantly being gathered in genomic research (e.g. there are at least 26 billion base pairs representing the various genomes available in the server of the National Center for Biotechnology Information (NCBI) (Zoheir Ezziane, 2006)). Cohen (2004) also pointed out the basic concepts in molecular cell biology, and outlined the nature of the existing data, and illustrated the algorithms needed to understand cell behavior. Methods from bioinformatics and computational biology are increasingly used to augment or leverage traditional laboratory and observation-based biology. These methods have become critical in biology due to recent changes in our ability and determination to acquire massive biological datasets, and due to the ubiquitous, successful biological insights that have come from the exploitation of those data. This transformation from a data-poor to a data-rich field began with DNA sequence data, but is now occurring in many other areas of biology such as mining hypermethylated genes in breast cancer tumor subtypes (Amin et al., 2012; Zoheir Ezziane, 2006; Smolinski et al., 2008a,b).

Recently, we have seen a new era of computational intelligence emerging that is focussing on the principles, theoretical aspects, and design methodology of algorithms gleaned from nature. Examples include artificial neural networks inspired by mammalian neural systems, evolutionary computation inspired by natural selection in biology, simulated annealing inspired by thermodynamics principles, and swarm intelligence inspired by collective behavior of insects or micro-organisms interacting locally with their environment causing coherent functional global patterns to emerge. Arenas et al. (2011) reviewed the adaptations of graphical processing unit (GPU) in scientific computing applications, particularly in the fields of computational biology and bioinformatics.

CI is a well-established paradigm with current systems having many of the characteristics of biological computers and capable of performing a variety of tasks that are difficult to do using conventional techniques. It is a methodology involving adaptive mechanisms and/or an ability to learn that facilitate intelligent behavior in complex and changing environments, such that the system is perceived to possess one or more attributes of reason, such as generalization, discovery, association and abstraction. These methodologies including technologies such as neural networks, fuzzy systems, rough sets, evolutionary computation, swarm intelligence, probabilistic reasoning and multi-agent systems. Recent trends aim to integration of different components to take advantage of complementary features and to develop synergistic systems leading to hybrid architectures such as neuro-fuzzy systems, genetic-fuzzy systems, evolutionary-fuzzy systems, evolutionary-neural networks, genetic programming neural networks or rough-neural, fuzzy-rough approaches for problem solving (Ritchie et al., 2007; Weyde and Dalinghaus, 2004; Hassanien and Slezak, 2006; Hassanien, 2007; Papageorgiou and Froelich, 2012).

The objective of this article is to present to the CI and bioinformatics research communities some of the state-of-the-art in CI applications to bioinformatics and motivate research in new trend-setting directions. Hence, we review and discuss some representative methods to provide inspiring examples to illustrate how CI techniques can be applied to solve bioinformatics problems and how bioinformatics data can be analyzed, processed, and characterized by CI. We want to stress that the literature in this domain is of course huge and that therefore the work that we include here is only exemplatory and focussing on recent advances, while many other interesting research approaches had to be omitted due to space limitations such as k-means clustering, decision trees, and case-based reasoning (Horng et al., 2009).

The rest of the article is organized as follows. Section 2 introduces the fundamental aspects of the key components of modern computational intelligence including including neural networks, restricted Boltzmann machine, deep belief network, fuzzy logic, rough sets, evolutionary algorithms (EA), genetic algorithms (GA), swarm intelligence, artificial immune systems and support vector machines. Sections 3 and 4 review some of the work that utilize computational intelligence for gene expression data clustering and classification respectively. An overview of work in computational intelligence based approaches in protein sequence classification is given in Section 5. Section 6 reviews and discusses some successful approaches to illustrate how CI can be applied to gene selection problems. CI in DNA fragment assembly and in multiple sequence alignment is discussed in Sections 7 and 8 respectively. Protein structure prediction is reviewed in Section 9. Challenges and future trends are presented in Section 10 and an extensive bibliography is provided.

2. Computational intelligence

In this section, we present an overview of modern CI techniques including artificial immune systems, neural networks, restricted Boltzmann machine, deep belief network, fuzzy logic, rough sets, EAs, GAs, swarm intelligence, and SVMs,

2.1. Artificial neural networks

Artificial neural networks (ANN) have been developed as generalizations of mathematical models of biological nervous systems.