Parallel Computing 42 (2015) 60–74

Contents lists available at ScienceDirect

Parallel Computing

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

Geometrical motifs search in proteins: A parallel approach

Marco Ferretti [1], Mirto Musci *

Dipartimento di Ingegneria Industriale e dell'Informazione, University of Pavia, Via Ferrata 3, 27100 Pavia, Italy

ARTICLE INFO

Article history:
Available online 12 October 2014

Keywords:
Proteins
Secondary structure
Motif extraction
Hough transform
Data parallelism
OpenMP

ABSTRACT

The analysis of the 3D structures of proteins is a very important problem in life sciences, since the geometric set-up of proteins has a deep relevance in many biological processes. The complexity of the analysis and the continuous increase in the number of proteins whose 3D structure is known, call for efficient and quick algorithms. Parallel processing is becoming an enabling tool for such research. A key component in the geometric description of a protein is the structural motif, a 3D element which appears in a variety of molecules and is usually made of just a few simpler structures, the secondary structures elements (SSEs).

This paper is an extended version of Ferretti and Musci (2013), and presents the Cross Motif Search (CMS) and the Complete CMS (CCMS) algorithms, two highly optimized and efficient parallel methods to detect the presence and location of all common motifs of secondary structures in a given protein pair (CMS) or across an arbitrary large dataset of proteins (CCMS). The analysis builds on existing approaches, such as Secondary Structure Co-Occurrences (SSC), based on the General Hough Transform (GHT) technique. The main difference between our proposal and the state of the art is the innovative focus that CMS puts on the geometric description of the structural motifs, which could be simply viewed as vectors in a 3D space, rather than on the topological/biological description employed by competing algorithms, such as ProSMoS, PROMOTIF or MASS. The advantage of a geometrical approach is that it enables to retrieve the exact location of the common substructures in a protein pair.

The paper analyzes all possible forms of serial and parallelism optimization of the proposed algorithms, both shared memory and message passing. It introduces a complete parallel implementation of CMS, based on OpenMP, and discusses its scalability on shared-memory architectures. Both small-scale and medium-scale testing shows that the methods produces very interesting results in real applications, and scales nicely up to the eight-processor limit. More in-depth testing also shows that the scalability limit is due to the inner structure of the problem, and that the similarities among proteins and the chosen tolerance for the analysis highly affect the overall performance.

1. Introduction

The analysis of the 3D structures of proteins is a very important problem in life sciences, since the geometric set-up of proteins has a deep relevance in many biological processes. The complexity of the analysis and the continuous increase in

* Corresponding author.
E-mail addresses: marco.ferretti@unipv.it (M. Ferretti), mirto.musci01@universitadipavia.it (M. Musci).
[1] Principal corresponding author.

http://dx.doi.org/10.1016/j.parco.2014.09.007
0167-8191/

the number of proteins whose 3D structure is known, call for efficient and quick algorithms. Parallel processing is becoming an enabling tool for such research. A key component in the geometric description of a protein is the structural motif, a 3D element which appears in a variety of molecules and is usually made of just a few simpler structures, the secondary structures elements (SSEs).

This paper presents the Cross Motif Search (CMS) and the Complete CMS (CCMS) algorithms, two highly optimized and efficient parallel methods to detect the presence and location of all common motifs of secondary structures in a given protein pair (CMS) or across an arbitrary large dataset of proteins (CCMS). The analysis builds on existing approaches, such as Secondary Structure Co-Occurrences (SSC) [2,3] based on the General Hough Transform (GHT) technique [4]. The main difference between our proposal and the state of the art is the innovative focus that CMS puts on the geometric description of the structural motifs, which could be simply viewed as vectors in a 3D space, rather than on the topological/biological description employed by competing algorithms, such as ProSMoS [5,6], PROMOTIF [7] or MASS [8]. The advantage of a geometrical approach is that it enable to retrieve the exact location of the common substructures in a protein pair, with varying degrees of tolerance. With respect to other geometrical algorithm, such as SSM [9], CMS is way more precise, as SSM only gives a similarity score for the geometrical structures of pair of proteins. As we will show, our main goal is to look for previously unknown common geometrical structures in so-called "unfamiliar" proteins. The main shortcoming of our method with respect to the state of the art is performance, however, as precise geometrical approach are inherently slower. Thus, an efficient implementation becomes even more important.

The paper analyzes all possible forms of serial and parallelism optimization of the proposed algorithms. It introduces a complete parallel implementation of CMS, based on OpenMP, and discusses its scalability on shared-memory architectures. Both small-scale and medium-scale testing shows that the method produces very interesting results in real applications, and scales nicely up to the eight-processor limit. More in-depth testing also shows that the scalability limit is due to the inner structure of the problem, and that the similarities among proteins and the chosen tolerance for the analysis greatly impact the overall performance.

The paper is organized as follows: Section 2 discusses the basics of the application background and the terminology for describing proteins in terms of their secondary structure elements; Section 3 introduces the Hough transform approach to geometrical motif identification and details our proposed algorithms and their computational complexity; Section 4 analyzes the possible optimizations of the serial versions of the algorithms, including a greedy version of CMS that reduces computation time up to two order of magnitudes; Section 5 examines the sources of parallelism in the application, both shared memory and message-passing and thoroughly described our OpenMP parallel implementation. In Section 6 we present a performance study of our implementation, and showcase some application results. Section 7 concludes the paper and introduces our next research project, that is a full-scale implementation on a massively parallel systems with a mixed OpenMP/MPI approach.

This paper is an extended version of Ferretti and Musci [1]. Generally speaking the previous contribution was mostly concerned about an old version of the algorithm, the Entire Motif Search, which is briefly treated here in Section 3.3. In more detail, this paper presents the following differences with respect to Ferretti and Musci [1]: (1) the application background is presented more clearly and in more details, for the benefit of he reader; (2) two new algorithms are introduced and discussed, namely Cross Motif Search (CMS) and Complete Cross Motif Search (CCMS); (3) the complexity analysis has been extended to include both the CMS and the CCMS case; (4) more serial optimizations are discussed, including the extremely efficient greedy variant of CMS; (5) the parallel implementation is discussed in much more details; (6) the testing suite has been vastly extended, to present a more sound assessment of the parallel performance of the algorithm; the focus has obviously been shifted from the old EMS algorithm to CMS.

2. Application background

2.1. Protein structure: a hierarchy of descriptions

Proteins are large and complex polymers, and scientists are accustomed to describing their structure at several levels, in order to better understand their spatial arrangement and behavior.

1. The **primary structure** describes a protein as the sequence of all the residues (amino-acids) composing its polypeptide(s).
2. The **secondary structure** describes a protein in terms of its **secondary structure elements** or **SSEs**.
3. The **tertiary structure** describes the overall 3D shape of a single polypeptide, comprising the atoms it is made of and their exact position.
4. For proteins that are made up of more than one polypeptide, the spatial arrangement of the polypeptides (also called protein subunits) is referred to as **quaternary structure**.

For the purpose of our work, we are mainly interested in the secondary structure. A secondary structure element is a small segment of the protein chain which can be identified as a recurring pattern among a multitude of proteins. Each SSE is made of multiple residues shaped in a peculiar 3D structure.

The first two and the most commonly recurring SSEs are the α-helix and the β-sheet as described by Pauling et al. [10].