CHARLES UNIVERSITY PRAGUE
faculty of mathematics and physics

To:

The Board of Doctoral Study
Faculty of Mathematics and Physics
Charles University in Prague

July 19th, 2010

Report on the PhD thesis by David Hoksza

The topic of the thesis – similarity search in protein databases – concerns an interdisciplinary research area including database systems (content-based search techniques) on one side, and computational biology subfield (proteomics) on the other. As the interdisciplinary nature of the problem spans over biology and computer science, it requires an significant effort to understand the fundamentals of both disciplines and to recognize opportunities where they could be synergistically combined. One of the basic tasks in computational biology (or bioinformatics) is to search for proteins with similar biological function. This could be accomplished on the sequence level (proteins represented as strings of letters representing amino acids) or on the structural level (proteins represented as 3D shapes). The thesis contributes to effective (accurate) and efficient (fast) protein similarity search at both representation levels. In Chapter 1 the reader is briefly introduced into the fundamentals of proteomics, while Chapter 2 overviews the state-of-the-art approaches to access methods for proteins modeled on the sequential level. The Chapter 3 consists from the first contribution of the thesis, which is an experimental study on utilization of metric indexing methods for speeding the retrieval of protein sequences. In Chapter 4 another speedup method is presented for the sequence-based techniques, but this time focused on direct optimization of the sequence similarity measure based on dynamic programming. In Chapter 5 the candidate overviews the current approaches to search in databases of proteins represented by structures. In Chapter 6 the candidate proposes his third contribution, the nonalignment-based model for structural similarity retrieval of proteins. The fourth contribution, as presented in Chapter 7, is an alignment-based structural model, showing superior accuracy of protein classification over present solutions. Finally, in Chapter 8 a possibility of indexing the model proposed in Chapter 7 in metric space is discussed.

The thesis demonstrates a comprehensive insight of the candidate into the domain, while the proposed contributions to the research area are significant and original. The candidate also proved his ability to independently recognize, formulate and develop novel approaches to protein similarity search. Because the focus of the thesis is wide enough, thorough, and presents the area of proteomics from the computer-science point of view, which is a novel self-contained presentation of the research field, I recommend to publish an extension of the thesis as a monograph on database methods in proteomics.

The results presented in this thesis have been published in proceedings on a number of representative international bioinformatic and database conferences (IEEE, Springer, ACM) and also in an international bioinformatic journal. The candidate also participated on an international summer school abroad, oriented to computational biology.

I recommend the candidate to obtain the PhD degree.

Doc. RNDr. Tomáš Skopal, Ph.D.
supervisor