An open-source framework for evaluating MS/MS score functions: "pluggable scoring" in X! TANDEM

Brendan MacLean², Jimmy Eng¹, Ronald C. Beavis¹, Martin McIntosh¹
¹Computational Proteomics Laboratory, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
²LabKey Software, LLC, Seattle, WA, USA

Overview
The X! TANDEM software tool has been adapted to allow novel score functions to be "plugged in". The goal is to facilitate and foster the development, implementation, and evaluation of better ways of identifying peptides by tandem mass spectrometry (TANDEM) search. This is achieved by modifying TANDEM such that core infrastructure is separated from the actual score function module, enabling the scoring class to be overridden. We present a general introduction to the software modifications that enable pluggable scoring, supply example score function modules, and discuss plans for future work.

Introduction
Peptide and protein identification via tandem mass spectrometry and sequence database searching has become ubiquitous in the realm of proteomics. As the field of proteomics grows and attracts a larger computational audience, there is an ever increasing emphasis on developing more accurate and sensitive computational methods to identify the constituent proteins in a sample. Most of this research has focused on MS/MS database searching and more specifically on how to better score a match between a query spectrum and candidate peptides from a sequence database. The most prevalent MS/MS database search tools in use today are commercial tools that are typically proprietary and not amenable to user modification. The recent availability of open source MS/MS search tools such as TANDEM, OMSSA², and ProBioID allows researchers the opportunity to understand the mechanisms behind how they work and make modifications to them as needed. For this project, we focused on adding pluggable scoring to the TANDEM software as we found that the software is well-architected, contains many rich features, has rapidly gained traction within the proteomics community, and allows us to be compatible with and take advantage of the rich and growing set of tools of the GPM project. By extending TANDEM to make the score function pluggable, we hope to lower the barrier to development and evaluation of new scoring algorithms with the goal of fostering research in the realm of mass spectrometry-based identifications.

Methods
The frameworks for various MS/MS database search tools are conceptually very similar. Most involve a three component process: reading in input spectra and database sequence databases, identifying candidate peptides of the same mass as the input spectra, comparing sequence against spectrum to generate a score, storing relevant peptide hits as the database is being searched, and, lastly, writing out the search results. So the implementation and testing of a new score function typically requires developing an entire search tool just to investigate one (albeit very important) subcomponent.

To facilitate the development and evaluation of novel score functions, we modified the TANDEM program to allow scoring modules to be easily pluggable. The modifications enable adding new (and modifying existing) score functions without any modifications to the rest of the source code.

The modifications to the TANDEM source involve first separating out general scoring infrastructure from the actual native TANDEM score algorithm. Infrastructure for scoring functions, such as a state machine for choosing which peptide sequences to score, was left in the base class. Specific scoring decisions were moved into functions that were then made virtual so they could be overridden by other modules. These virtual functions, which represent the true "Pluggable Scoring API", are marked within the mscore.h file in the TANDEM project (see insert to the right).

The source code modifications also include the creation of a generalized plug-in manager. The plug-in manager maps a plug-in type to a set of named plug-ins, such as mscore.h and mscore.cpp in the score function directory for details. The creation of a new score plug-in involves implementation of a single C++ source file and simple corresponding header file. Scoring modules can be added or removed simply by adding or removing the pair of .cpp./h files from the project and recompiling. No other changes are required, either to the Makefile or the other source files, to add/remove these scoring plug-ins.

The modifications made to TANDEM, as well as TANDEM's native hyposcore score function (referred to as the mscore function) are available as part of the default TANDEM distribution. The T-score module, which is a score function implementing simple spectral processing + scoring and intended to be used as a template for developing new plug-ins, along with other current and future scoring modules, will be available from the CPL website at proteomics.fhcrc.org pending publication.

Conclusions
• We implemented pluggable scoring by modifying TANDEM's score search program to facilitate the addition of new scoring algorithms without requiring modifications to the rest of the TANDEM source code.
• We present example score function plug-in modules that can be used as templates to develop new scoring algorithms.
• We demonstrate a framework for evaluating/comparing the performance of the various plug-ins within the CPAS analysis system.

Future work includes plans to investigate the development of more accurate and sensitive scoring algorithms; hopefully this work encourages other computational researchers to do the same.

This work was funded by NCI contract 23XS144A

References