Tree Genetics & Genomes (2009) 5:723–727
DOI 10.1007/s11295-009-0220-2

SHORT COMMUNICATION

Spruce proteome DB: a resource for conifer proteomics research

Dustin Lippert · Mack Yuen · Jörg Bohlmann

Received: 28 February 2009 / Revised: 28 February 2009 / Accepted: 27 May 2009 / Published online: 24 June 2009

Abstract Proteomics research is hampered in many organisms due to a lack of an appropriate reference genome sequence that can be used in the interpretation of tandem mass spectrometry data for the identification of proteins. Public DNA sequence repositories have grown to considerable size and can, in most cases, serve to provide at least partial interpretation of a large-scale proteomics dataset. However, when species-specific sequences or sequences from a closely related species are available, a boutique sequence database can provide considerable increases in specificity, confidence, and completeness of protein identification. Here, we describe the development of a protein database from a large-scale expressed sequence tag and full-length complementary DNA sequencing project in the economically and ecologically important spruce (*Picea*) genus.

Keywords *Picea* · Proteome · Database

Introduction

In recent years, there has been growing appreciation of the need to apply systems biology approaches that go beyond the genome level to the study of plant science (Cui et al. 2008; Long et al. 2008; Nelson et al. 2008). This is due to the realization that the assignment of gene function and the understanding of dynamic molecular phenotypes depend greatly on the ability to measure changes occurring beyond the level of gene expression. One method of achieving this is through the measurement and characterization of the proteins being expressed within a biological system (i.e., proteomics). Proteomics represents a rapidly developing technical discipline that encompasses a wide range of activities such as the analysis of changing protein abundance (Bachi and Bonaldi 2008), posttranslational modifications (de la Fuente van Bentem et al. 2008), and functional protein interaction networks (Collura and Boissy 2007). However, proteomics methods continue to be underutilized in the area of plant biology outside of their application in well-defined model systems like *Arabidopsis thaliana* and *Oryza sativa* (Chen and Harmon 2006; Pan et al. 2005). As a rapidly developing discipline, it is through the creation of new tools that the value of these methods will be unlocked in other plant systems.

Proteomics research relies heavily on the use of tandem mass spectrometry, and an average dataset typically consists of tens to hundreds of thousands of individual mass spectra. By extension, proteomics research is critically dependent upon the availability of sequence databases for the rapid and unsupervised interpretation of these spectra to provide meaningful peptide sequence assignments and the associated protein identifications. Organisms for which the genome has not been sequenced have typically been at a disadvantage with respect to the practical application of proteomics methods. These organisms typically rely on searching against sequences from related species that share sequence identity with the organism under study. For species of spruce (*Picea* spp.) and other conifers, the most closely

Communicated by J. Dean

D. Lippert · M. Yuen · J. Bohlmann (✉)
Michael Smith Laboratories, University of British Columbia,
#321-2185 East Mall,
Vancouver, British Columbia V6T 1Z4, Canada
e-mail: bohlmann@msl.ubc.ca

Present Address:
D. Lippert
Manitoba Center for Proteomics and Systems Biology,
The University of Manitoba,
Winnipeg, Manitoba, Canada

related genomes are all from evolutionarily distant angiosperm species (e.g., *A. thaliana*, rice, poplar, grapevine). However, it has been shown that distantly related sequences function poorly in the interpretation of proteomics data (Huang et al. 2006). The spruce proteome database (DB) described here was assembled from the sequence data produced during a large-scale expressed sequence tag (EST) and full-length complementary DNA (FLcDNA) sequencing project in spruce (Ralph et al. 2008) with representative sequences taken from *Picea sitchensis* (Sitka spruce), *Picea glauca* (white spruce), and *Picea glauca* × *engelmannii* (interior spruce). Spruce proteome DB is an expansion of the databases used in prior proteomics studies performed in these conifer species (Lippert et al. 2005, 2007, 2009) and consists of a set of related protein databases representing these three spruce species and hybrids studied in the Treenomix project (www.treenomix.ca). Spruce proteome DB is, to our knowledge, the most comprehensive and appropriate sequence resource for studying conifer and other gymnosperm proteomes. Spruce proteome DB complements other database resources that provide general information on conifers (e.g., The Gymnosperm Database; http://www.conifers.org/index.html) and conifer genomics (e.g., TreeGenes, http://dendrome.ucdavis.edu/treegenes/).

Database construction

Spruce proteome DB was constructed according to the pipeline illustrated in Fig. 1. The final database was assembled using a set of 437,705 ESTs from Sitka spruce (168,424), white spruce (242,931), and interior spruce (26,350) in addition to 10,579 FLcDNA sequences from Sitka spruce (Pavy et al. 2005; Ralph et al. 2006, 2008). EST sequences from each species were clustered independently using parallel contig assembly program (Huang et al. 2003), resulting in a separate set of contigs and singletons for each of the three species. The FLcDNAs were used as is. All nucleotide sequences were compared sequentially to the *Arabidopsis* protein database followed by all the plant sequences in NCBInr using BLASTx (Altschul et al. 1990). Matches were accepted with *e* values less than 1×10^{-5} and the annotations were associated with the relevant query sequences. *Arabidopsis* annotations were chosen preferentially over NCBInr annotations when both were obtained. Sequences that were not matched using BLAST were subsequently analyzed using GENEMARK-E (Besemer and Borodovsky 2005; Borodovsky et al. 2003). GENEMARK-E is an ab initio gene finder that attempts to identify potential genes in eukaryotic sequences that have no known homolog. In spruce, roughly 10% of the ESTs fall into this category. The fact that these sequences were obtained from cDNAs (i.e., transcribed genes) suggests that they may represent proteins that are unique to spruce and possibly other conifers or gymnosperms. To this end, GENEMARK-E was used in order to uncover putative open reading frames (ORFs) within these sequences. Matches were labeled as Genemark XX YY, where XX was the position of the ORF start site and YY was the position of the stop site within the original EST sequence. These entries were trimmed so as to only contain the sequence of the putative ORF. All remaining unmatched EST and contig sequences were translated in six frames and each frame was then included in the final build of spruce proteome DB as a separate entry derived from that nucleotide sequence annotated as "Hypothetical Protein based on EST sequence." The structure of the sequence annotations is shown in Fig. 2. The description line for each sequence contains up to four different pieces of information, depending on the outcome of the annotation process for that sequence. These consist of a short string representing the species of origin, a number unique to the contig or EST sequence, the annotation obtained from BLASTx or the GENEMARK-E algorithm, and the expect value of any BLAST match that was found.

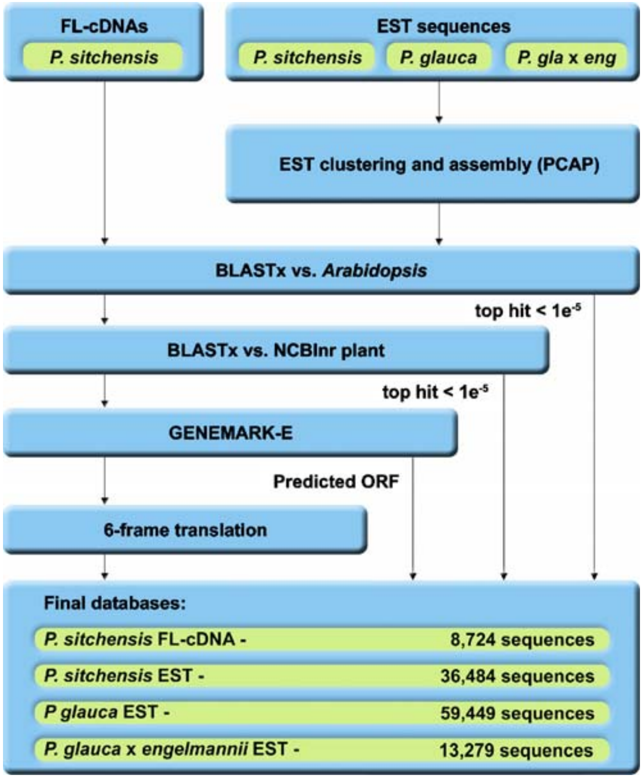

Fig. 1 Flowchart illustrating the series of steps performed during the construction of spruce proteome DB. Sequence inputs are listed as are the output subset databases. The size of each database is indicated by the number of individual entries contained within each sequence set

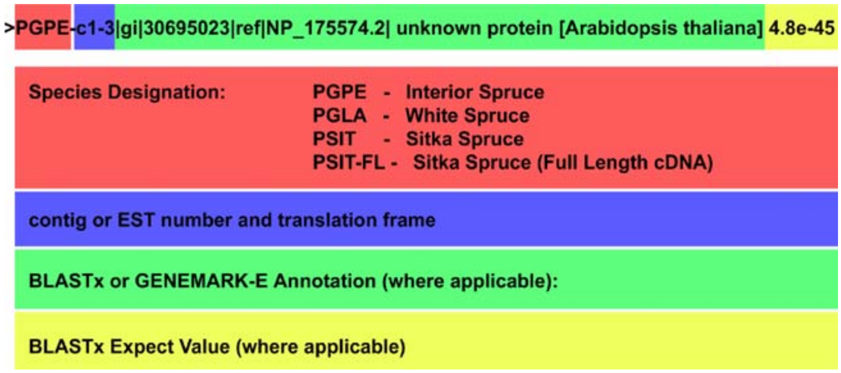
>PGPE-c1-3|gi|30695023|ref|NP_175574.2| unknown protein [Arabidopsis thaliana] 4.8e-45
Species Designation:
PGPE - Interior Spruce
PGLA - White Spruce
PSIT - Sitka Spruce
PSIT-FL - Sitka Spruce (Full Length cDNA)
contig or EST number and translation frame
BLASTx or GENEMARK-E Annotation (where applicable):
BLASTx Expect Value (where applicable)

Fig. 2 The information contained in the definition line for each database entry describes the source of the entry. An example is shown here and explanatory notes are listed for each component of the definition line. The species from which the sequence was taken is indicated at the beginning of the line. This is followed by a set of symbols that indicate whether the sequence is a contig, EST, or full-length cDNA and the reading frame of the original nucleotide sequence that is represented in the entry. Finally, the result of any BLASTx-derived annotation or GENEMARK-E result is listed along with an expect value if appropriate

In addition to the main databases described above, decoy versions of each database have been produced that contain head-to-tail reversed sequences. These decoy databases are provided separately and can be combined with spruce proteome DB to assess the level of false-positive protein identification that is obtained from any proteomics dataset following a database search. The implementation of this approach for the analysis of proteomics data has been previously described (Huttlin et al. 2007). In brief, matches to reverse sequences represent random incorrect matches and the scores that are obtained against these reverse sequences can be used to empirically determine an appropriate score cutoff when interpreting the result of a proteomics database search.

Database implementation and access

Spruce proteome DB can be accessed and used for the interpretation of tandem mass spectrometry data through an instance of the global proteome machine (GPM; Craig and Beavis 2004) at http://treenomix.ca/Home/ResearchActivities/FunctionalGenomics/ProteinProfiling/SpruceDB.aspx (username: tggreview; password: treenomix). Users can upload their peak extracted data in any GPM compatible format (e.g., .mgf, .mzxml). The complete spruce proteome DB can also be obtained by direct download for use with other proteomics analysis software platforms. The database is provided in fasta format and should be compatible with all commercial and open-source platforms. At present, the database has been successfully tested with both Mascot (Perkins et al. 1999) and ProteinPilot (Applied Biosystems, Foster City, CA, USA) as alternative search engines.

Database performance

The utility of spruce proteome DB was assessed by comparing it against the set of plant protein sequences available in the NCBI database for the interpretation of peptide tandem mass spectrometry data. The data used were taken from a previously completed study performed in Norway spruce (*Picea abies*) and represents the liquid chromatography–tandem mass spectrometry analysis of a set of 22 cation exchange fractions from the separation of a protein digest. A more detailed description of the Norway spruce protein sample source and method of preparation can be found elsewhere (Lippert et al. 2009). Since Norway spruce is not well represented within the available EST resources, the dataset was searched against the Sitka spruce portion of spruce proteome DB, including both the EST and

Table 1 Performance comparison of spruce proteome DB and NCBInr plant for Norway spruce protein identification from tandem mass spectrometry data

	Sitka spruce proteome DB	Plant translated UniGene (NCBI)	Sitka spruce UniGene (NCBI)
Number of protein sequences in the database	45,208	3,492,768	100,601
Number of spectra submitted	19,345	19,345	19,345
Number of proteins identified (log(*e*)≤−3.0)	492	393	407
Minimum log(*e*)	−196.7	−128.4	−145.7

FLcDNA sequences. For comparison, the same data were analyzed using the same search criteria against two different publicly available sequence databases. One search was performed against the plant translated UniGene sequences from NCBI and the second was performed specifically against the translated Sitka spruce UniGene sequences contained within the larger NCBI database. Spruce proteome DB contains 47,208 unique Sitka spruce protein sequences as compared with 3,492,795 translated plant UniGene sequences from the NCBI database and 100,601 translated Sitka spruce UniGene sequences (Table 1). The searches were performed using default parameters for the type of mass spectrometer used for data collection, and a brief summary of the results has been tabulated. With spruce proteome DB, 492 proteins were identified with a log(expect) value less than or equal to −3.0 from the test dataset. In comparison, 393 proteins were identified using the much larger NCBI database and 407 proteins were identified when using only the Sitka spruce sequences from the NCBI database. There was also a corresponding increase in the confidence with which these proteins were identified in spruce proteome DB as indicated by the log(expect) values of the most confidently identified protein (−196.7 for spruce proteome DB vs. −128.4 using NCBI and −145.7 using the Sitka spruce sequences from NCBI). This result reflects an increase in both the specificity for individual peptide matches in addition to an increase in the number of peptides identified per protein. There is a clear benefit to the use of spruce proteome DB for the interpretation of proteomics data even for related species not directly represented in the database itself.

Conclusion

In its present form, spruce proteome DB provides a resource tailored to the analysis of proteomics data in species of spruce, which is one of the largest species groups in the conifers including many of the economically and ecologically most important forest tree species of the northern hemisphere. This database may also provide benefit in the analysis of proteomics data from other conifers and gymnosperms. Future versions of this database will expand upon the depth of spruce proteome coverage as new sequencing efforts are undertaken but will also attempt to gather and process sequences from other conifer and gymnosperm species to expanding both the size, diversity of species included, and the general utility of the database for the analysis of gymnosperm proteomics.

Acknowledgements We would like to thank Richard Varhol and An He (Canada's Michael Smith Genome Sciences Center, Vancouver, British Columbia) for bioinformatics support. The work described in this paper was supported with a grant from the Natural Sciences and Engineering Research Council of Canada (to J.B.) and funding from Genome British Columbia and Genome Canada for the Treenomix Conifer Forest Health project (www.treenomix.ca; to J.B.). J.B. is a University of British Columbia Distinguished University Scholar.

References

Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Bachi A, Bonaldi T (2008) Quantitative proteomics as a new piece of the systems biology puzzle. J Proteomics 71:357–367

Besemer J, Borodovsky M (2005) Genemark: Web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33:W451–W454

Borodovsky M, Lomsadze A, Ivanov N et al (2003) Eukaryotic gene prediction using genemark.hmm. Curr Protoc Bioinformatics 4:4.6

Chen S, Harmon AC (2006) Advances in plant proteomics. Proteomics 6:5504–5516

Collura V, Boissy G (2007) From protein–protein complexes to interactomics. Subcell Biochem 43:135–183

Craig R, PBeavis RC Cortens J (2004) Open source system for analyzing, validating, and storing protein identification data. J Proteome Res 3:1234–1242

Cui J, Li P, Li G et al (2008) Atpid: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. Nucleic Acids Res 36:D999–1008

de la Fuente van Bentem S, Mentzen WI, de la Fuente A et al (2008) Towards functional phosphoproteomics by mapping differential phosphorylation events in signaling networks. Proteomics 8:4453–4465

Huang X, Wang J, Aluru S et al (2003) Pcap: a whole-genome assembly program. Genome Res 13:2164–2170

Huang M, Chen T, Chan Z (2006) An evaluation for cross-species proteomics research by publicly available expressed sequence tag database search using tandem mass spectral data. Rapid Commun Mass Spectrom 20:2635–2640

Huttlin EL, Hegeman AD, Harms AC et al (2007) Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. J Proteome Res 6:392–398

Lippert D, Zhuang J, Ralph S et al (2005) Proteome analysis of early somatic embryogenesis in *Picea glauca*. Proteomics 5:461–473

Lippert D, Chowrira S, Ralph SG et al (2007) Conifer defense against insects: proteome analysis of Sitka spruce (*Picea sitchensis*) bark induced by mechanical wounding or feeding by white pine weevils (*Pissodes strobi*). Proteomics 7:248–270

Lippert DN, Ralph SG, Phillips M et al (2009) Quantitative itraq proteome and comparative transcriptome analysis of elicitor-induced Norway spruce (*Picea abies*) cells reveals elements of calcium signaling in the early conifer defense response. Proteomics 9:350–367

Long TA, Brady SM, Benfey PN (2008) Systems approaches to identifying gene regulatory networks in plants. Annu Rev Cell Dev Biol 24:81–103

Nelson T, Gandotra N, Tausta SL (2008) Plant cell types: reporting and sampling with new technologies. Curr Opin Plant Biol 11:567–573

Pan S, Carter CJ, Raikhel NV (2005) Understanding protein trafficking in plant cells through proteomics. Expert Rev Proteomics 2:781–792

Pavy N, Paule C, Parsons L et al (2005) Generation, annotation, analysis and database integration of 16, 500 white spruce est clusters. BMC Genomics 6:144

Perkins DN, Pappin DJ, Creasy DM et al (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567

Ralph SG, Yueh H, Friedmann M et al (2006) Conifer defence against insects: microarray gene expression profiling of Sitka spruce (*Picea sitchensis*) induced by mechanical wounding or feeding by spruce budworms (*Choristoneura occidentalis*) or white pine weevils (*Pissodes strobi*) reveals large-scale changes of the host transcriptome. Plant Cell Environ 29:1545–1570

Ralph SG, Chun HJ, Kolosova N et al (2008) A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). BMC Genomics 9:484