Applying Bioinformatic Techniques to Identify Cold-adaptive Genes in Oats

A Masters Dissertation by
Henrik Thorburn

Supervised by Björn Olsson

Environment

- Large amount of data
- Largely unstructured data
- No standard method for gene identification
- Important to make use of already acquired knowledge and data
- Easy access to proper tools, such as BLAST, FASTA, PROSITE and HMMER

Hypothesis

- It is possible to identify cold-adaptive genes in the *Avena sativa* genome by looking for close homologues to already known cold-adaptive genes in related organisms.
- Possible to do so with the help of standard bioinformatic tools.

Prerequisite

It is assumed that it is possible to look into organisms that are related to the organism being researched, search for homologues, and predict genes based on the homologies.

The chance that a newly sequenced gene has homologues that are already analysed is above 70 %

(Bork et al. 1998)

Aim

Investigate whether standard bioinformatic tools, applied to EST data, can be used to identify genes with a certain functionality

- Develop a rapid prediction method
- Apply the method and predict cold-genes
- Estimate the accuracy of the method

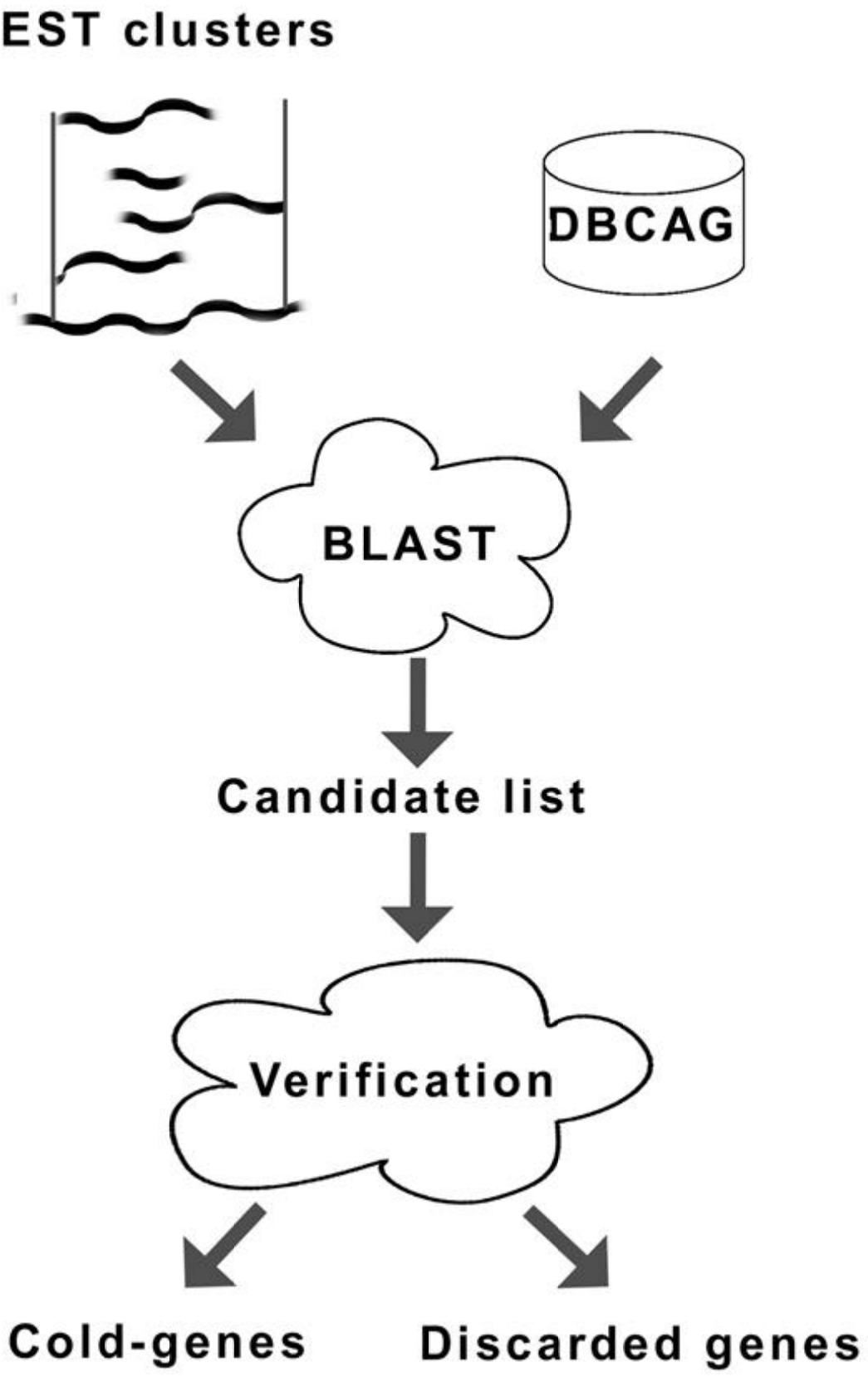
Limitations

- Only look for similarities in the *Avena sativa* EST data compared to other plants

- Only use the selected tools
  - BLAST
  - PROSITE
  - HMMER

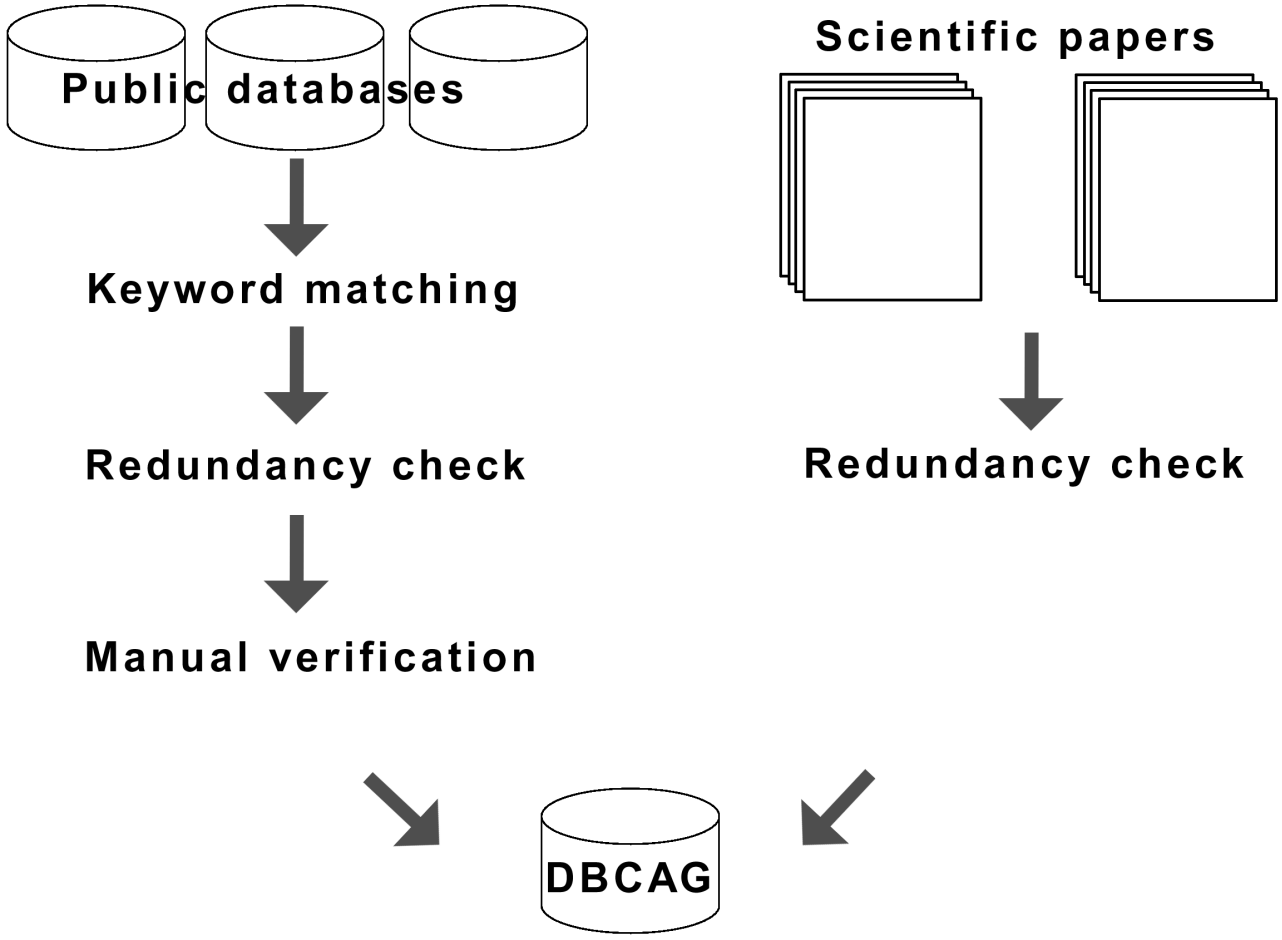
Method description

- Three method steps
  - Create a database containing genes with known cold-acclimation properties
  - Identify cold-gene candidates in the *Avena sativa* EST dataset
  - Verify each candidate to:
    - Estimate the accuracy of the method
    - Prune the candidate-list


EST clusters
DBCAG
BLAST
Candidate list
Verification
Cold-genes
Discarded genes

Step 1 : DBCAG

(DataBase of Cold-Acclimation Genes)



Step 2 : Identify cold-gene candidates in *Avena sativa*

BLAST each gene in DBCAG against the *Avena sativa* EST dataset.

Step 3 : Verification of cold-gene candidates

Reason 1: Estimate the accuracy of the rapid gene prediction

Reason 2: Prune the candidate-list

Multiple tool prediction:

- Protein and Nucleotide BLAST search (GenBank)
- PROSITE pattern search (PROSITE)
- Profile HMM search (Pfam)

- Candidates:
  - BLAST hit with $E < 10^{-5}$ to DBGAC *sure*
- Likely cold-genes:
  - As above, and:
  - No BLAST hit to non-DBCAG *sure* gene with lower *E* than best BLAST hit to DBCAG *sure*
  - No PROSITE match to non-cold family or Pfam hit with $E < 10^{-5}$ to non-cold family unless a better cold-family matches as well
- Verified cold-genes:
  - As above, and:
  - Hit to Prosite or Pfam cold-family with $E < 10^{-5}$

    or
  - BLAST hit with $E < 10^{-40}$ to DBCAG *sure*, and
  - No BLAST hit to non-DBGAC *sure* with lower *E*

Related prediction tools

- GeneQuiz

  Thorough tool that predicts gene function in a way similar to the presented method. This allows comparative analysis

  (Hoersch et al. 2000, Andrade et al. 1999)

- ProtFun

  Assigns a protein or enzyme to pre-defined classes. Hard to use to predict and verify cold-acclimation genes. Difficult to compare results.

  (Jensen et al, 2001)

Results

- DBCAG
  - 72 genes from scientific papers
  - 304 additional putative cold-genes identified through keyword search and examined (214 were kept and 90 discarded)
  - DBCAG has three categories:
    - **sure** **157**
    - *unsure* *26*
    - *similar* *103*
  - In this work only DBCAG *sure* is used to identify and verify cold-gene candidates

Results cont.

- Candidate identification
  - 227 ESTs
  - 7 cold-gene candidates
- Candidate verification
  - 2 non cold-genes
  - 1 likely cold-gene
  - 4 verified cold-genes

Candidates A to G

Candidate	PROSITE	Pfam	BLAST	GeneQuiz
A (non cold-gene)			Unknown protein	Cytocrome
B (verified)	ADH	ADH	ADH2, ADH3	ADH
C (verified)			RLT, BLT, TACR	Low temperature induced protein
D (likely)		LEA (weak)	COR, COR14a, WCS19	COR14a
E (verified)			Cold-regulated PEA-NMT*	PEA-NMT*
F (non cold-gene)			BAC	PM29
G (verified)	PS50020		WCS19	COR14a

* Phospho-ethanolamine n-methyltransferese

Analysis

- Verified and likely cold-genes (B,C,D,E,G)
  - Consensus between tools
  - Consensus with GeneQuiz
- Non cold-genes
  - Weak consensus between tools
  - Weak consensus with GeneQuiz

Observation:

ESTs with atleast one high-scoring BLAST hit are often by different tools predicted to have a similar function.

Avena sativa, preliminary results

- Candidate identification
  - 9513 ESTs (8775 after filtering out zero-length sequences)
  - 135 cold-gene candidates
- Candidate verification

35	unlikely	(26 %)
15	likely	(11 %)
85	verified	(63 %)

Conclusions

- It is possible to perform gene prediction in EST data with standard bioinformatic tools
- Consensus between prediction tools as well as with GeneQuiz, possibly due to:
  - Good reliability
  - Assignment of gene and protein function based on homologies
- The method is simple to reuse, and also to modify to look for other functions

Future work

- More advanced scoring systems
- Automatic gene categorisation
- Thorough comparison of tool prediction consensus
- Enhancing the database of genes known to possess the sought functionality