

Exhaustive assembly and functional classification of publically available plant ESTs

Andrea Hansen, Christine ME Schueller and Jean Hani

Biomax Informatics AG, Lochhamer Straße 11, D-82152 Martinsried, Germany



Biomax Informatics AG
Lochhamer Str. 11
D-82152 Martinsried
Tel. +49 89 895574-0
Fax. +49 89 895574-825

www.biomax.com

Is it necessary to undertake whole genome sequencing to describe the metabolome in plants? To address this important question we have analyzed all plant expressed sequence tags (ESTs) from *Arabidopsis thaliana*, *Oryza sativa* and *Hordeum vulgare* taken from the dbEST database (NCBI), clustered and assembled them, mapped the assemblies onto the Arabidopsis proteome (MATDB, <http://mips.gsf.de/proj/thal/db/>), transferred functional and Enzyme Commission (EC) number classifications and analyzed the resulting distributions in terms of biochemical pathway coverage. For the different steps of this analysis we used several Biomax software products, including the HarvESTer™ EST Clustering and Assembly System (Figure 1), the Pedant-Pro™ Sequence Analysis Suite and the BioXMT™ Pathway Analysis Tool.

The fully automatic analysis of HarvESTer EST Clustering and Assembly System includes several preprocessing steps (e.g., vector clipping, repeat masking, poly (A) tail removal) followed by clustering and assembly of the ESTs. The Clusterer module quickly subdivides EST data sets into batches based on sequence similarity. The Assembler module aligns the resulting clusters and creates a multiple alignment with a consensus sequence. Detailed statistics of the process are shown in Table 1.

The analysis of 177,934 EST sequences from *A. thaliana* resulted in 38,039 assemblies. From 339,025 *H. vulgare* ESTs, 62,211 assemblies were produced, and from 129,592 *O. sativa* ESTs, 41,433 assemblies were generated. The detailed sets of resulting clusters and assemblies are shown in Figure 2. The increasing number of contigs (data sets with more than one member) between the Clusterer and the Assembler may be due to splicing variants or gene families (see also differences between *O. sativa* and *H. vulgare*).

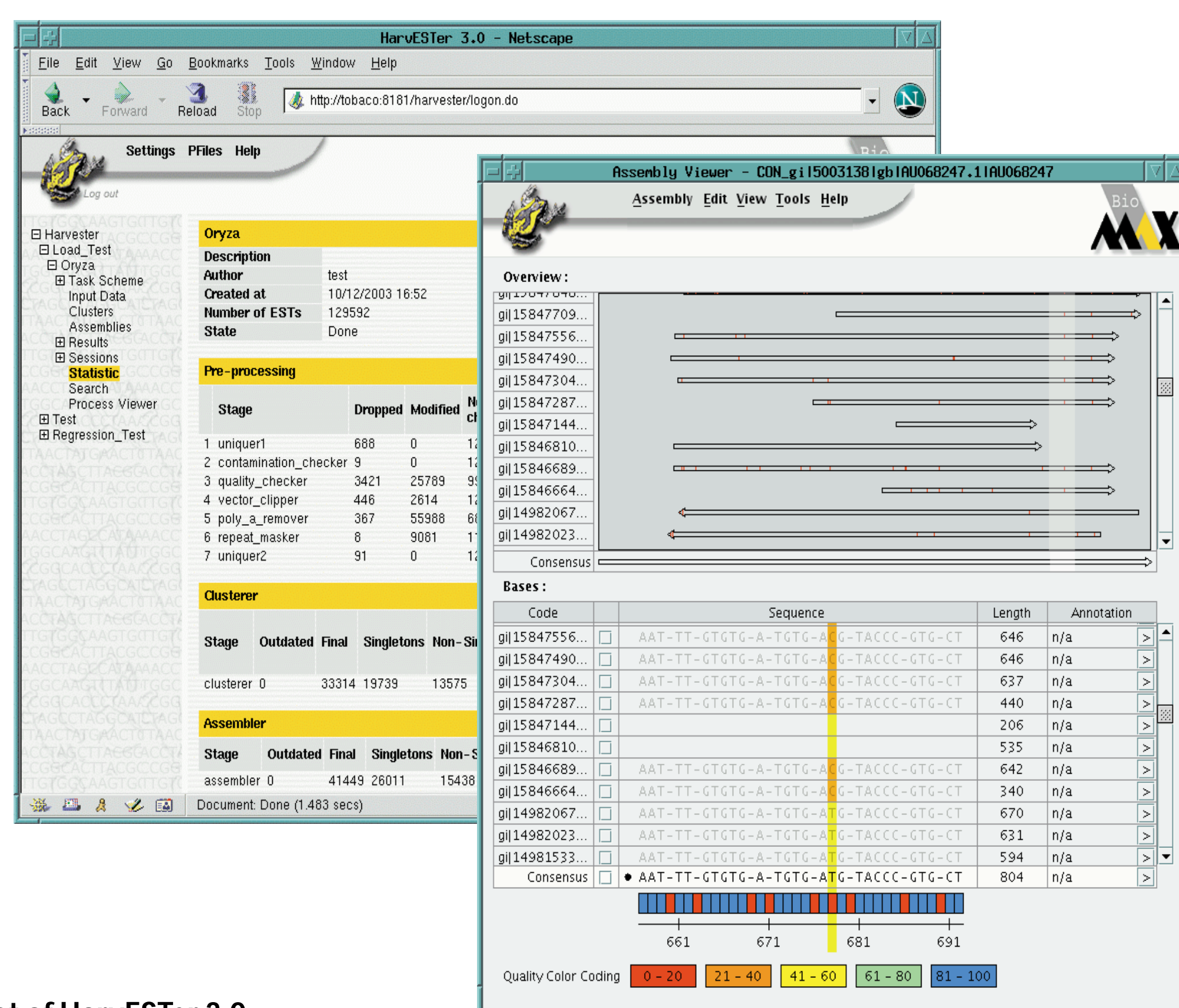


Figure 1: Screenshot of HarvESTer 3.0

The HarvESTer Clustering and Assembly System is a sophisticated high-throughput clustering and assembly tool for processing EST data. The left screenshot displays the statistical overview of *Oryza sativa*; the right screenshot shows a selected assembly, which may be edited afterwards.

The resulting EST consensus sequences were mapped to the proteome of *A. thaliana* (MATDB) using the Pedant-Pro™ Sequence Analysis Suite. For *A. thaliana*, 28,593 assemblies were mapped to the corresponding proteins using a BLAST cutoff score E-25. For *H. vulgare* and *O. sativa*, 29,805 and 10,644 assemblies, respectively, were mapped.

Functional assignments of the *A. thaliana* proteins based on the FunCat™ Functional Catalog were transferred to the corresponding EST consensus sequences. The FunCat catalog, developed at Biomax and MIPS*, is a hierarchically organized, controlled vocabulary that allows multidimensional, organism-independent annotation of protein properties. The catalog covers biological processes, molecular function and localization.

Comparison of the results of the plant EST sets was performed according to the functional distribution in the plants. Figure 3 displays the distribution of sequences that can be mapped to functional categories; Figure 4 shows the functional distribution within each organism. Analogously to the functional category mapping, Enzyme Commission (EC) numbers were assigned (Figure 5).

To determine which metabolic pathways are partially or completely found in the data sets, we have used the BioXMT™ Pathway Analysis Tool. For *H. vulgare*, 501 metabolic pathways were found containing at least one enzyme; 40% of these were complete (Table 2). For *O. sativa*, 480 pathways were found containing at least one enzyme; 38% of these were complete and for *A. thaliana* 482 pathways were found; 40% of these were complete.

Table 1: HarvESTer statistics

This table displays the detailed results of each module in HarvESTer.

a. Preprocessing modules: the numbers show how many sequences were dropped by the particular module (except for Repeat Masker which shows the number of modified ESTs). Two numbers separated by a "/" give the number of dropped and modified sequences. The first HarvESTer module is the Uniquer which removes redundant sequences from the data set. In the following step, the Contamination Checker screens for contamination using a manually curated database of transposons of the specific plant. After removal of low-quality regions by the Quality Checker, the Vector Clipper eliminates vector sequences at the end of the ESTs (NCBI UniVec DB). To avoid clustering of sequences based on poly A tails or repeats, the last two modules of the preprocessing remove poly (A) tails (Poly A Remover) and mask low-complexity repeats (Repeat Masker).

a. Preprocessing modules

Module	<i>A. thaliana</i>	<i>O. sativa</i>	<i>H. vulgare</i>
Uniquer	499	688	7758
Contamination Checker	51	9	110
Quality Checker	3,329/35,295	3,421/25,835	3,258/27,116
Vector Clipper	47/148	448/2,614	809/1,218
PolyA Remover	18/2,447	37/2,55,979	129/23,868
Repeat Masker	5,405	579,089	18761

b. Processing modules

Module	<i>A. thaliana</i>	<i>O. sativa</i>	<i>H. vulgare</i>
Clusterer	31,448 (18,224)	33,199 (13,511)	33,839 (18,255)
Assembler	38,039 (19,971)	41,433 (15,406)	62,211 (26,154)

b. Processing modules: the numbers of clusters and assemblies generated are shown (number of contigs are given in brackets). The processing starts with the clustering of the sequences using the hashed position tree algorithm (HPT), which builds a position tree with all sequences and then matches each particular sequence to the tree (Heumann et al. 1996). The resulting clusters are assembled using the CAP3 algorithm (Huang et al. 1999). The complete analysis was done using a four processor Linux® system.

Figure 2: Distribution of ESTs in the resulting clusters and assemblies

The Clusterer (grey line) produced 13,511 contigs (data sets with more than one member) and 19,688 singletons for *O. sativa* (data sets with one member). As expected, the number of contigs and singletons increases during assembly because splice forms and gene families which are condensed in one cluster are separated in the following assembly step. Therefore, 15,406 contigs and 26,027 singletons result from the assembly (yellow line). The corresponding results for *H. vulgare* are 18,255 contigs and 15,584 singletons for the clustering and 26,154 contigs and 36,057 singletons for the assembly process. The bigger increase of contigs in the *H. vulgare* data set may indicate that the data is incomplete or that there is greater variability of gene families or splicing forms in this plant. Analogously, for *A. thaliana*, 31,448 clusters fall into 18,224 contigs and 13,224 singletons, and 38,039 assemblies are composed of 19,971 contigs and 18,068 singletons (not represented in the figure).

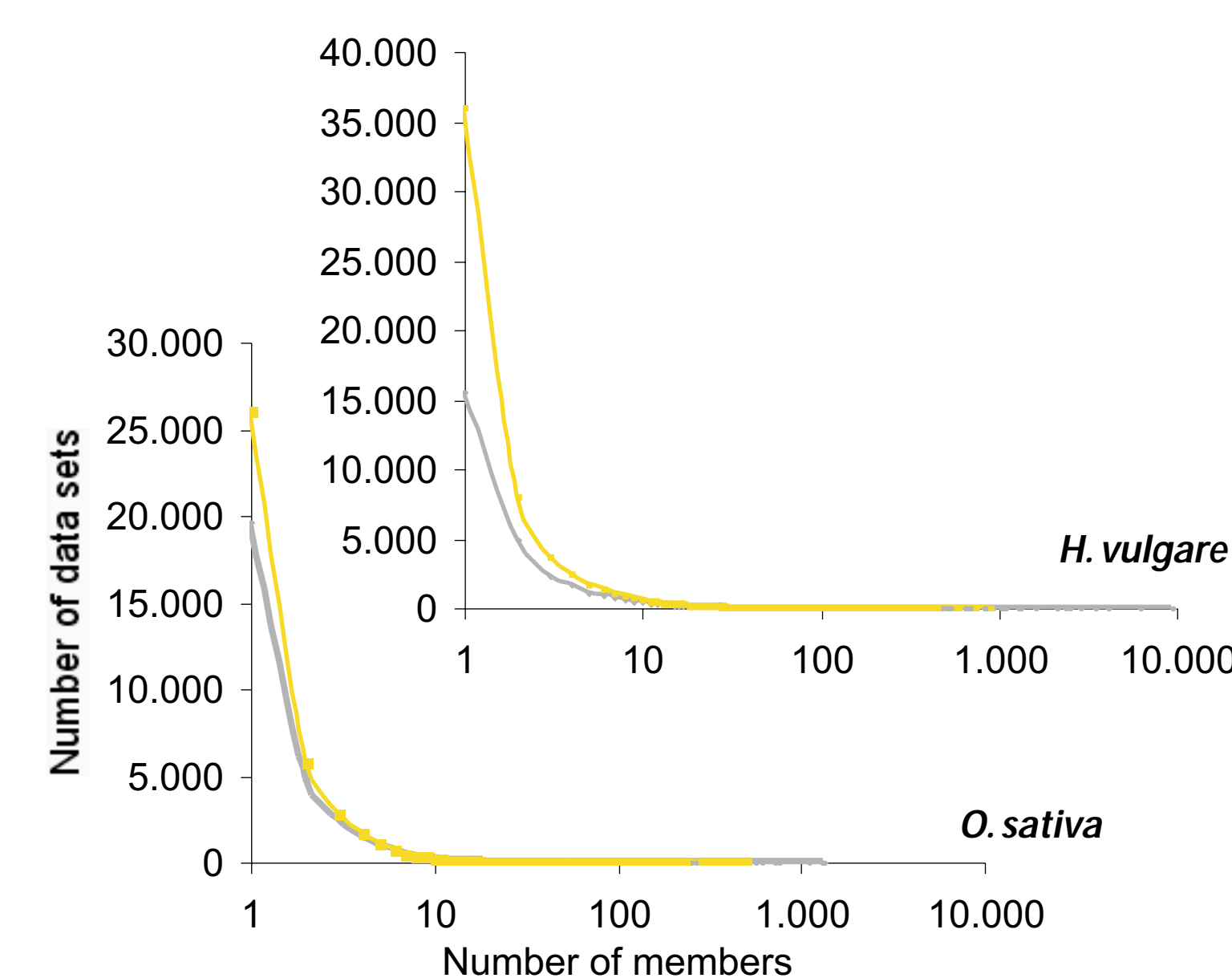


Figure 5: EC number distribution

5,937 assemblies of *H. vulgare* (shown in the figure) and 1,746 assemblies of *O. sativa* (not shown) can be mapped to known enzymes classified by the enzyme nomenclature system (EC numbers).

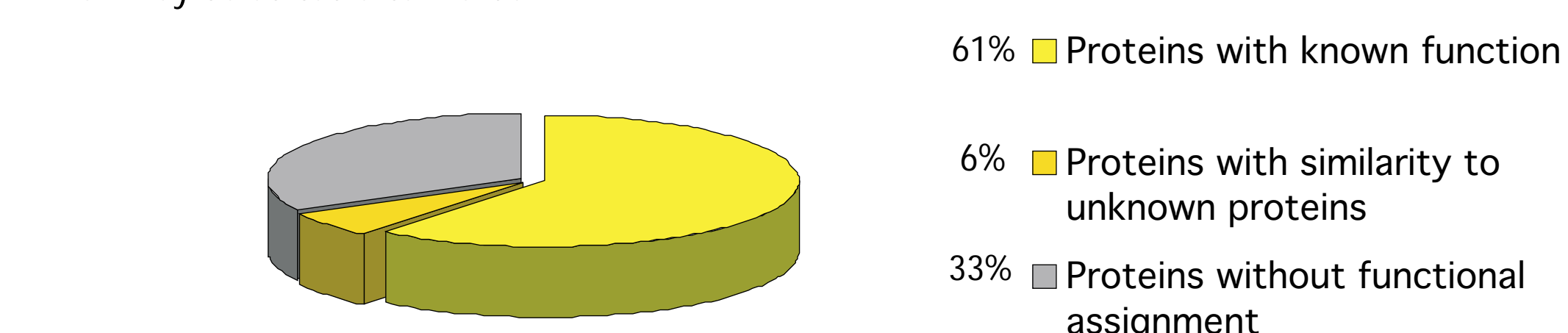
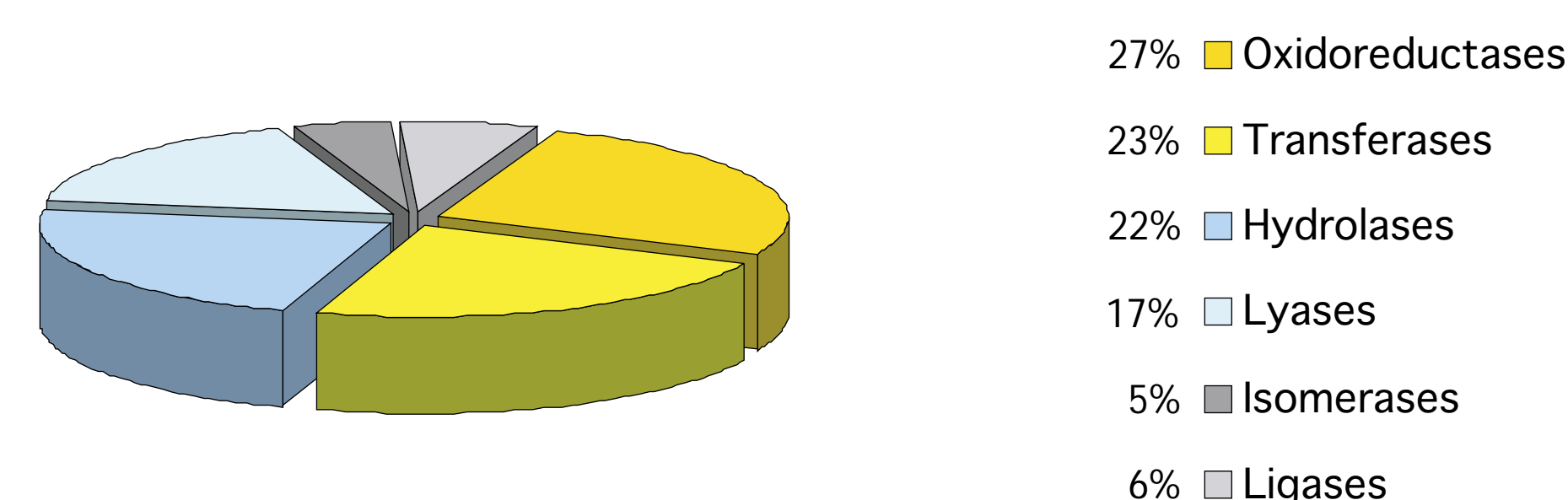


Figure 3: Distribution of all assemblies of *H. vulgare* which were mapped to the FunCat Catalog

Clustering and assembling the ESTs resulted in 62,211 assemblies. 29,805 of these can be mapped to a corresponding *A. thaliana* protein using a BLAST cut-off score of E-25. Among these matches, 61% can be assigned to proteins with known function, 6% match proteins with unknown function and 33% match proteins which are not yet functionally assigned in *A. thaliana*. The results for the *O. sativa* assemblies (41,433) are similar: 60% match proteins with known function, 6% match proteins with unknown function and 34% show similarity to proteins without functional assignment (not shown in the figure).

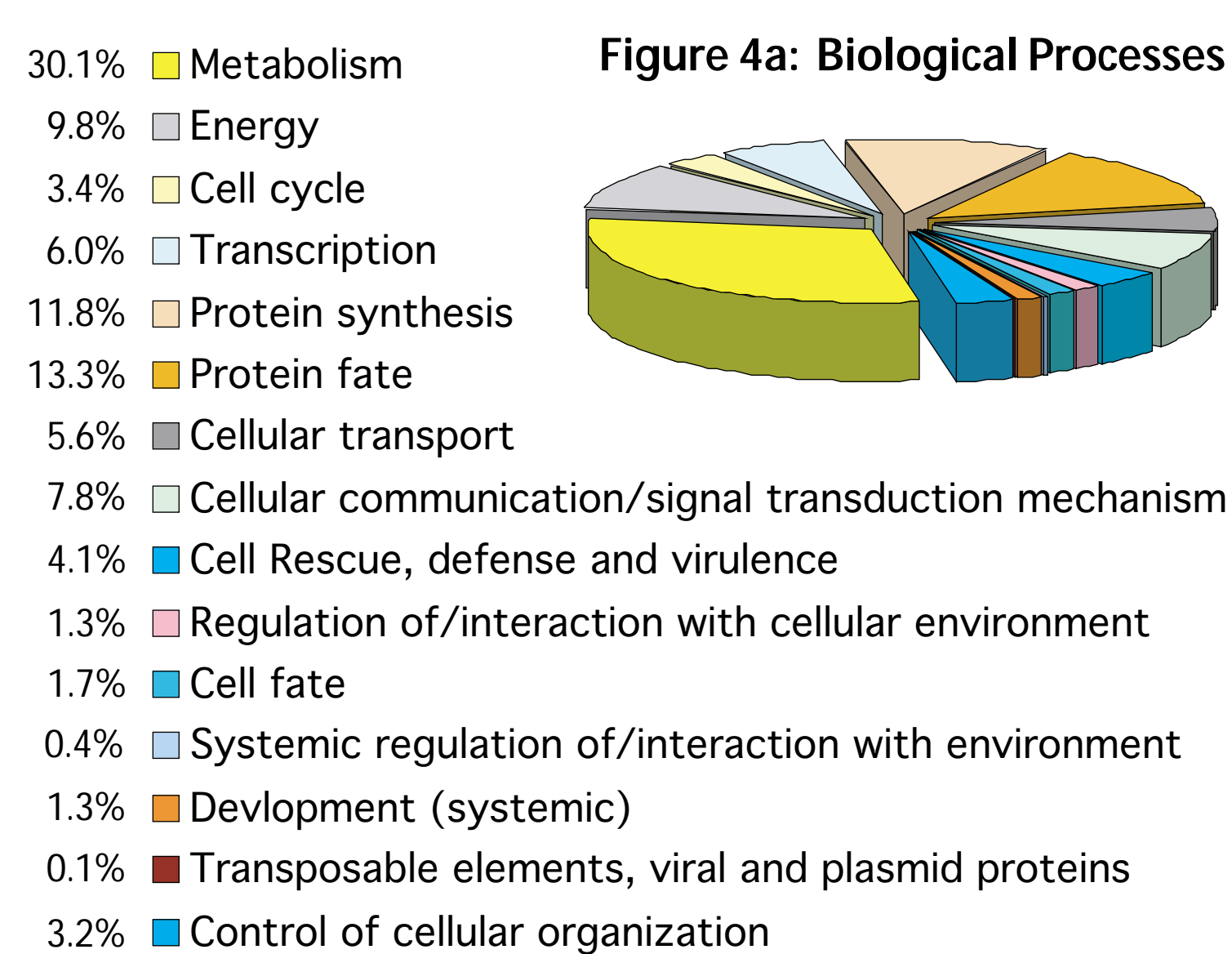


Figure 4a: Biological Processes

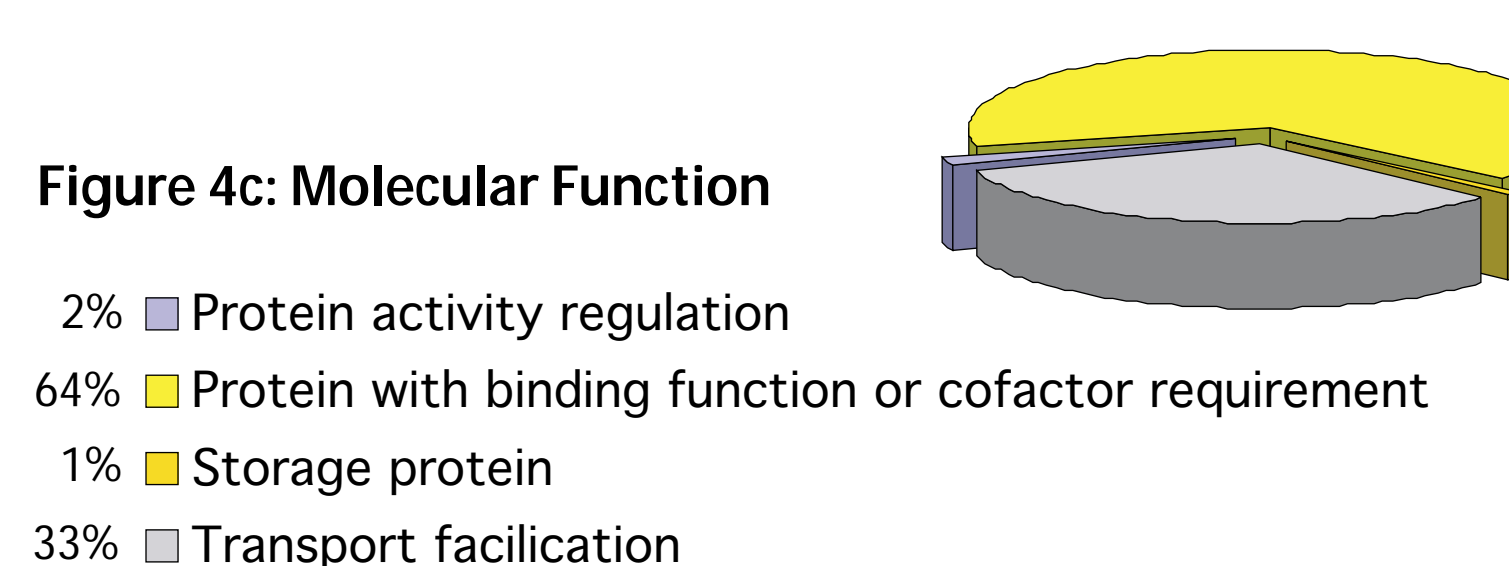


Figure 4c: Molecular Function

Figure 4b: Subcellular Localization

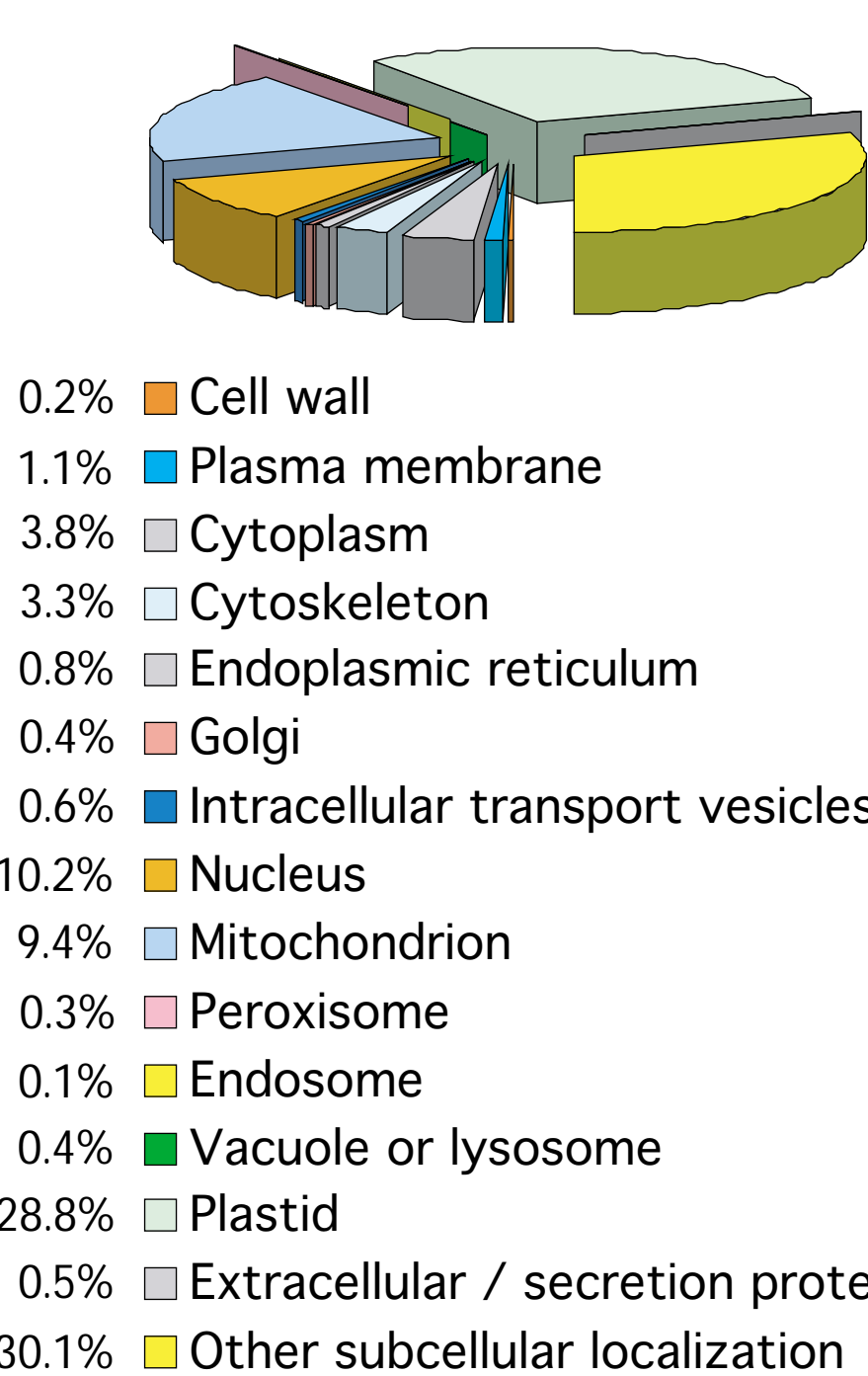


Figure 4: Detailed view of the functional assignments for *O. sativa* and *H. vulgare*.

For *O. sativa*, 13,787 sequences have been assigned by function. 6,336 of these were found to be involved in **Biological Processes** (Figure 4a). The pie graph shows the categories and percentages. For 3,885 assemblies a **Subcellular Localization** can be predicted (Figure 4b). For 1,731 sequences a **Molecular Function** can be predicted (Figure 4c). The remaining 1,342 assemblies have been mapped to unknown proteins. Analogously for *H. vulgare*, 38,728 functional predictions could be identified. 18,160 assemblies are involved in **Biological Processes** (distribution not shown). 11,118 consensus sequences have a predicted **Subcellular Localization** and 6,495 have a predicted **Molecular Function** (distribution not shown). 3,715 sequences have been mapped to unknown proteins.

References:
X. Huang et al (1999) *Genome Research* 9(9), 868-877

*The Munich Information Center for Protein Sequences (MIPS) is now the Institute for Bioinformatics. Biomax, HarvESTer, BioXMT, Pedant-Pro, and FunCat are registered trademarks of Biomax Informatics AG in Germany and other countries. Linux is a registered trademark of Linus Torvalds. Registered names, trademarks.

Table 2: BioXM Pathway Analysis Tool results

Mapped EC numbers (Figure 5) were used to identify the number of possible metabolic pathways in a specific organism. For each analyzed organism, the identified pathways were analyzed for completeness.

Pathway completeness (%)	<i>H. vulgare</i>		<i>O. sativa</i>		<i>A. thaliana</i>	
	Number of pathways	Percent of total pathways	Number of pathways	Percent of total pathways	Number of pathways	Percent of total pathways
100 - 90	206	40.4	185	38.5	194	40.2
89 - 80	8	1.6	6	1.3	8	1.7
79 - 70	10	1.9	5	1.0	6	1.2
69 - 60	55	10.8	34	7.0	40	8.3
59 - 50	82	16.1	72	15.0	79	16.4
49 - 40	24	4.7	29	6.0	24	5.0
39 - 30	51	10.0	65	13.5	56	11.6
29 - 20	42	8.2	52	10.8	41	8.5
19 - 10	24	4.7	18	3.8	24	5.0
9 - 1	8	1.6	14	2.9	10	2.1
Total	501	100	480	100	482	100

BIOINFORMATICS SOLUTIONS ... designed with you in mind