

Characterization and re-annotation of common genes found in 35 complete chloroplast genomes

Beatrice Kilel

School of Computational Sciences

George Mason University

Fairfax, VA

Motivation

- Many whole genomes currently available
- Annotation concerns on the completed genomes and annotation tools
- Whole genome comparisons not fully explored
- Knowledge gained from comparative genome analysis can be extrapolated across species

Scope Statement

- Re-annotation of the data whenever there are poor data and assign functions to new genes
- Gene Prediction
- Phylogenetic analyses using **Winclada** and **Nona** software on the complete chloroplast genomes

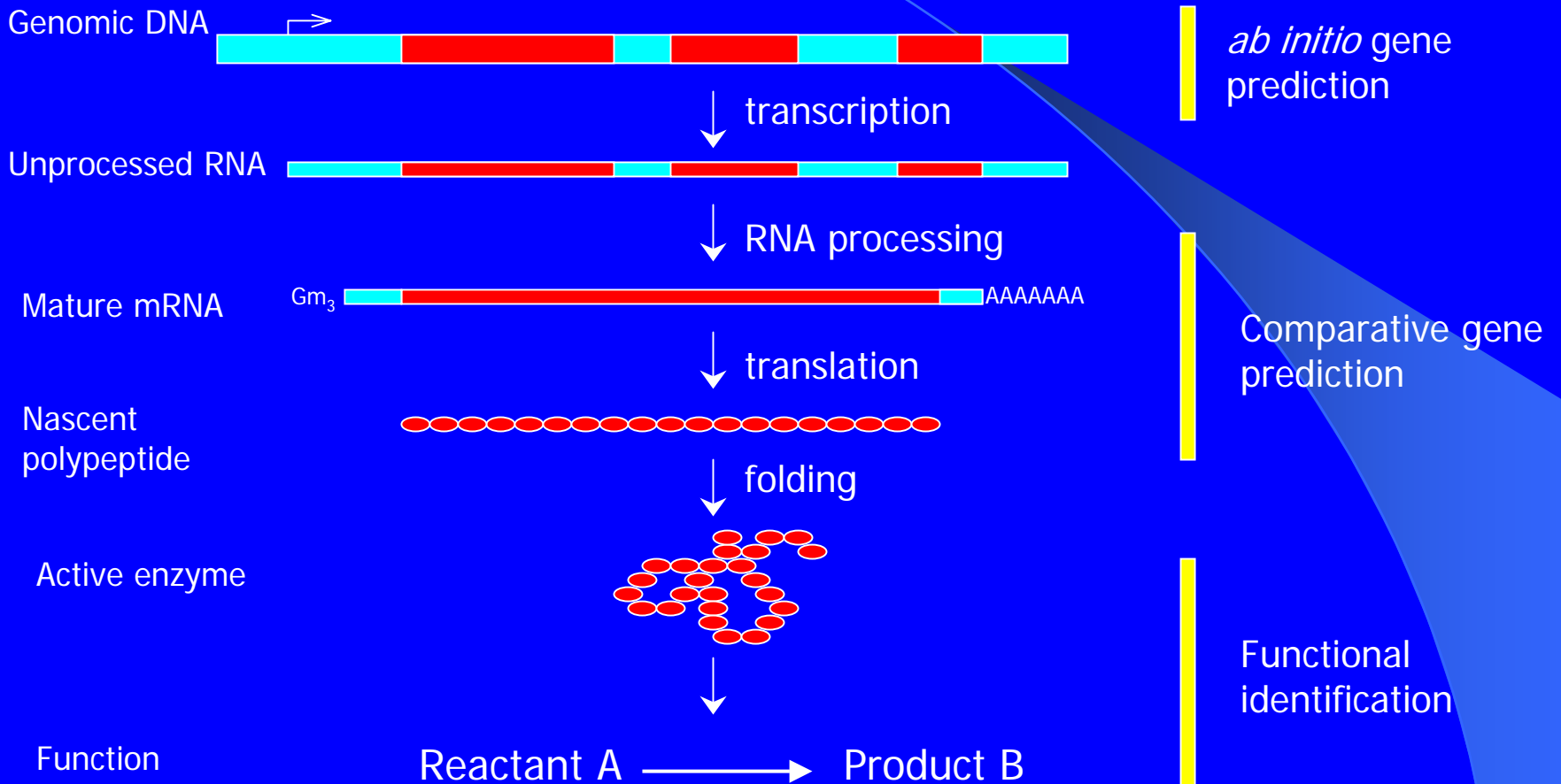
Why the chloroplast genome?

- Small size (~120 - 220 Kb, 120-150 genes), limited number of the repeated elements
- Well-conserved, low rate of mutations and hence excellent cladistic/phylogenetic tool
- Encode Proteins, rRNAs, tRNAs that are used in Photosynthesis (multifunctional organelle)

Why perform annotation?

- Obtain meaningful gene prediction
- Genome sequences are extremely large
- Need access to genome data both as a whole and in meaningful pieces
- Majority of the sequence in a genome doesn't correspond to known functionality

Fig 1. Annotation of eukaryotic genomes



Re-annotation

- The re-annotation process is essential in any sequence analysis for the review of the coding sequences, updating and citing of current data, postulating functions, and making name changes (Bocs et al. 2002)
- Manual review of data for concordance with transcript data, peptide similarity data as well as splice site usage (intron/exon boundaries)

Methods - Re-annotation

- Re-annotation to review genes in genome, update CDS, change functional classes, include current citations
- **GlimmerM** for re-annotation since it is trained for *Oryza sativa* and *arabidopsis thaliana*
- Results compared with **Genotator** automated annotation software for exon prediction by Genie
- **Artemis** annotation software for graphic displays in six frame translation
- **BlastP** for homology searches and gene prediction

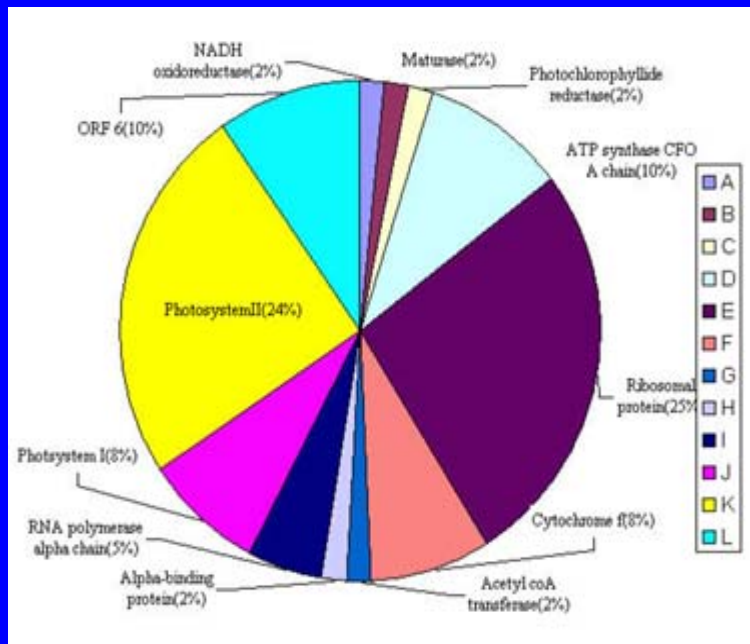
Re-annotation results

- *Triticum aestivum* originally had 18 protein encoding genes, 8 encoded stable RNA, after 4 more found to encode polypeptides
- Genes *rps16* and *chlL* absent in *Psilotum nudum* and present in *Adiantum capillus-veneris*
- Homologs of *Psilotum nudum orf83* or *orf119* not located in *Adiantum capillus-veneris*
- Drastic decrease could have resulted from frame-shifts and point mutations

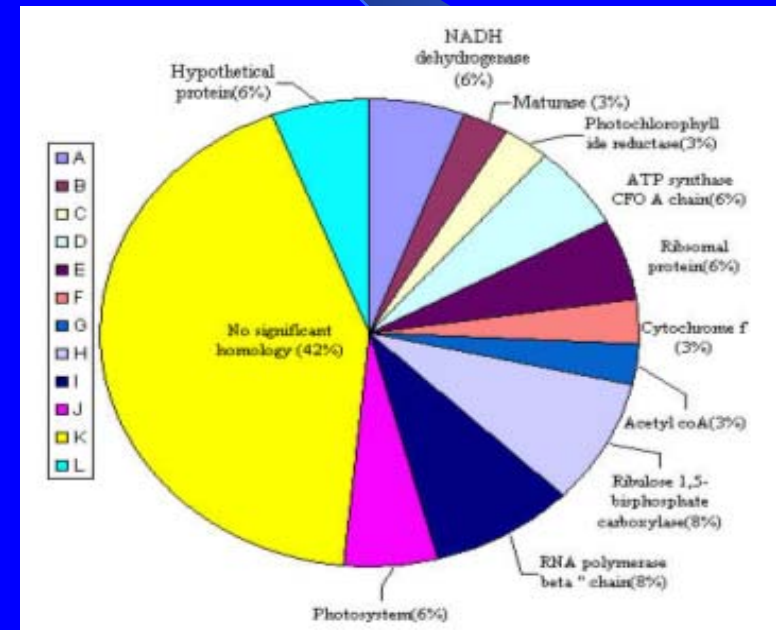
Table 1. Changes to protein-coding genes

Genome	CDS (Before)	CDS (After)	Genes/Kb	Coding%
<i>Adiantum capillus-veneris</i>	59	36	0.239	10.7
<i>Anthoceros formosae</i>	55	47	0.291	8.2
<i>Arabidopsis thaliana</i>	62	18	0.116	11.3
<i>Astasia longa</i>	27	22	0.229	10.5
<i>Atropa belladonna</i>	72	40	0.255	20.5
<i>Calycanthus fertilis</i> var. <i>ferax</i>	65	33	0.215	15.8
<i>Chaetosphaeridium globosum</i>	77	52	0.396	26.0
<i>Chlorella vulgaris</i>	70	62	0.405	21.4
<i>Cyanophora paradoxa</i>	85	60	0.442	26.1
<i>Cyanoideoschyzon merolae</i>	60	30	0.214	14.5
<i>Cyanothum caldarium</i>	83	62	0.375	25.2
<i>Epifagus virginiana</i>	31	12	0.171	22.9
<i>Engelena gracilis</i>	41	28	0.195	7.9
<i>Guillardia theta</i>	98	68	0.375	27.8
<i>Lotus japonicus</i>	60	48	0.318	21.5
<i>Marchantia polymorpha</i>	44	33	0.272	25.0
<i>Medicago truncatula</i>	Not in NCBI	22	0.177	20.5
<i>Mesostigma viride</i>	43	22	0.185	21.5
<i>Nephroselmis olivacea</i>	99	72	0.358	21.9
<i>Nicotiana tabacum</i>	57	37	0.237	18.8
<i>Odontella sinensis</i>	69	49	0.409	25.8
<i>Oenothera elata</i> subsp. <i>hookeri</i>	63	36	0.219	18.4
<i>Oryza sativa</i>	50	31	0.23	8.8
<i>Pinus koraiensis</i>	49	35	0.299	16.8
<i>Pinus thumbergii</i>	58	46	0.384	20.2
<i>Porphyra purpurea</i>	104	83	0.434	24.0
<i>Psilotum nudum</i>	53	36	0.259	18.7
<i>Spinacea oleracea</i>	69	49	0.325	19.8
<i>Toxoplasma gondii</i>	15	9	0.257	13.6
<i>Triticum aestivum</i>	75	33	0.245	10.6
<i>Zea mays</i>	33	18	0.128	8.3

Fig 2. Functional changes coding genes



Pre-annotation



Post-annotation

Adiantum capillus-veneris

Methods - Gene Prediction

- Masking known repeats and low complexity sequences using RepeatMasker
- Match to known genes
- Evidence from GlimmerM, Genscan
- Similarity to expressed sequences
- Comparative genomics
- Confirmation with molecular techniques
 - ** ideally, the blastn and blastx results should overlap - high interest feature

Gene Prediction results

- 5 functional groups: photosynthesis, metabolism, transport, transcription/translation, and protein kinases or phosphatases
- PSI, rubisco, ATPase may constitute an ancient core protein complex of most conserved genes

Gene Prediction ...

- hypothetical protein (GI:11465969) in *Nicotiana tabacum*, homologous to cemA- a heme-binding protein similar to ycf10 and ORF230 protein in *Oryza sativa* and *Zea mays*

Challenges

- Regions within a genome differ in gene density and GC content
- Statistical properties used in gene prediction methods can differ from genome to genome
- Evolution of function and sequence may not be as tightly linked as is sometimes believed
- Identification of gene families, orthologs, paralogs, xenologs

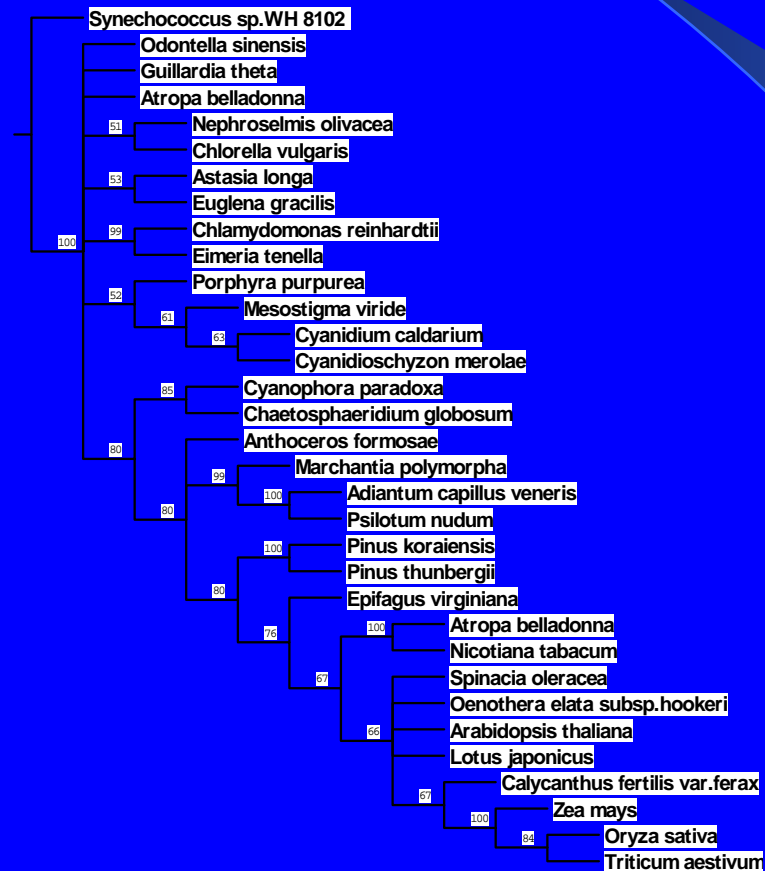
Comparative Analysis

- To infer relationships from proteins of known function to proteins of unknown function that are structurally similar
- When a relationship is not necessarily detectable from sequence comparison alone
- Gene predictions
- Explain the evolutionary distance between the species and function of genes (what and how) through the non-coding sequences

Methods - Phylogenetic analysis

- 19-gene data sets that are common were obtained from the GenBank
- **ClustalX** (Thompson et al. 1997) for complete sequence alignment- gap penalty (25 – 30), gap extension (6.66)
- **Winclada** shell (Nixon, 1999a) and **Nona** (Goloboff, 1994) for further analysis
- Jackknife analysis to test robustness of nodes of tree topology

Fig 3. Consensus of most parsimonious trees with Jackknife support values placed above the tree branches



Phylogenetic results

- Instances of local or large scale gene rearrangements were observed - can be used to explain species diversity
- Translocations, inversions, deletions, duplications
- Strong conservation of protein complexes essential for bioenergetics
- Clues on gene evolution and function from functionally linked protein networks on unknown ORFs

Fig 4. Gene Order (GeneOrder3.0)

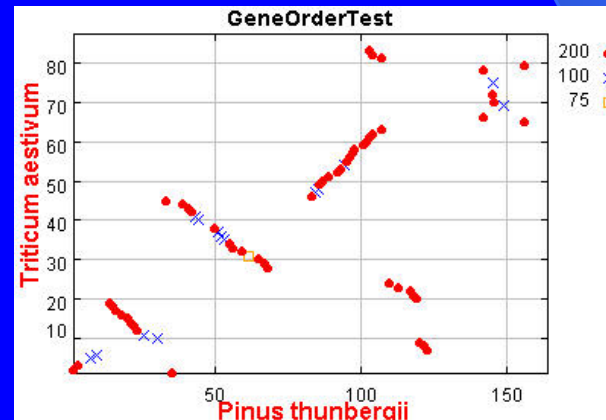
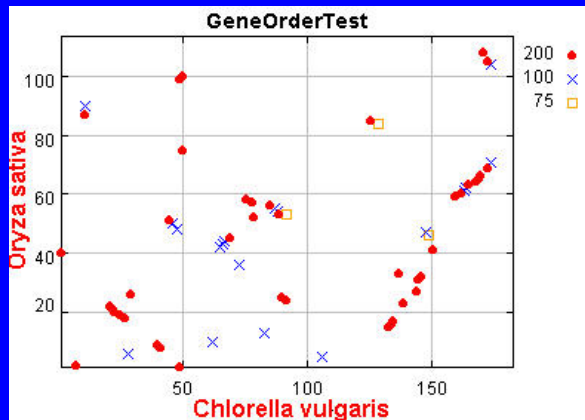
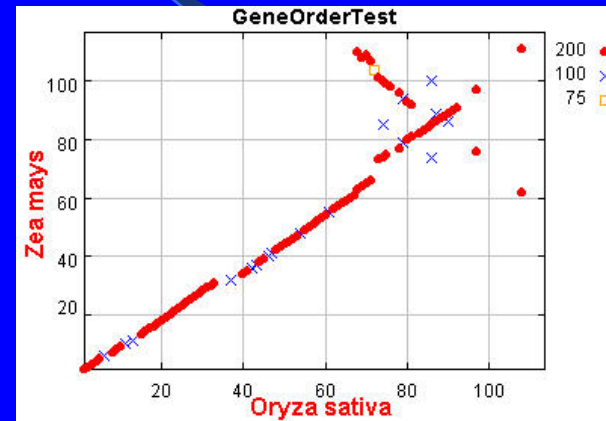
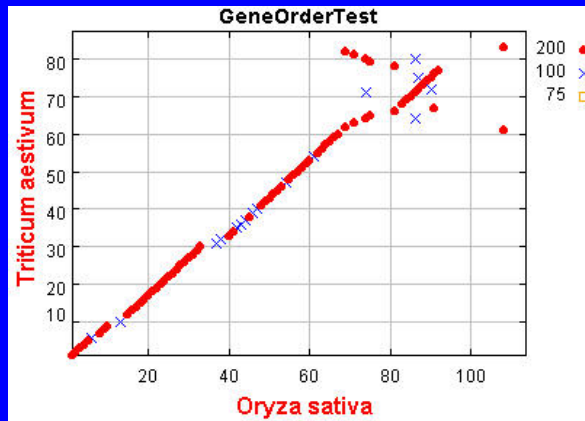


Fig 5. Network and PictTree of common genes found in chloroplast genomes

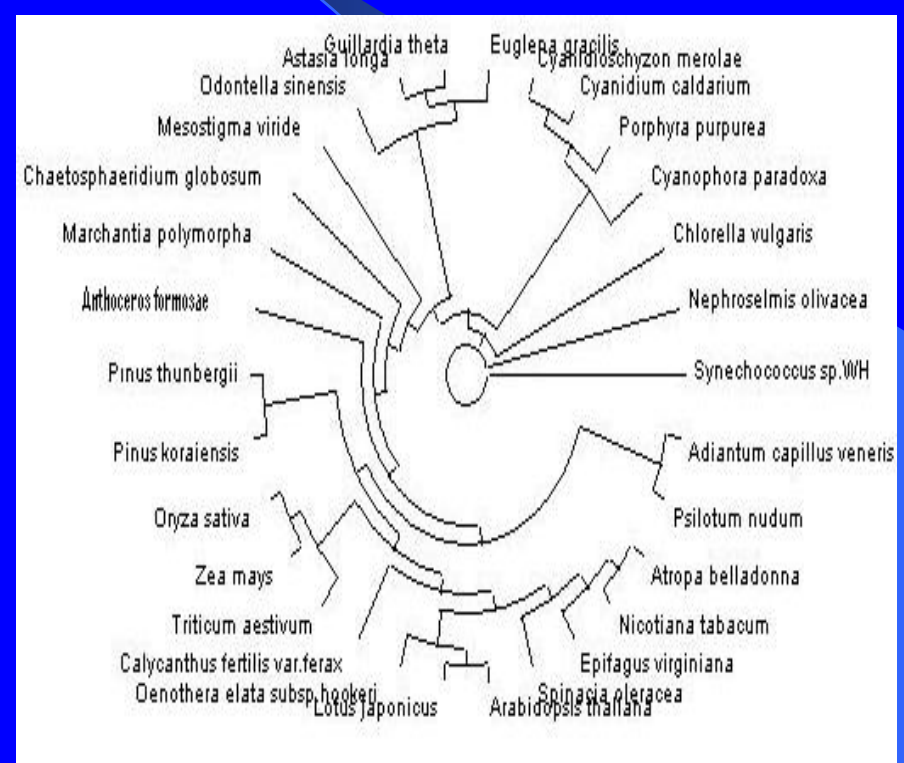
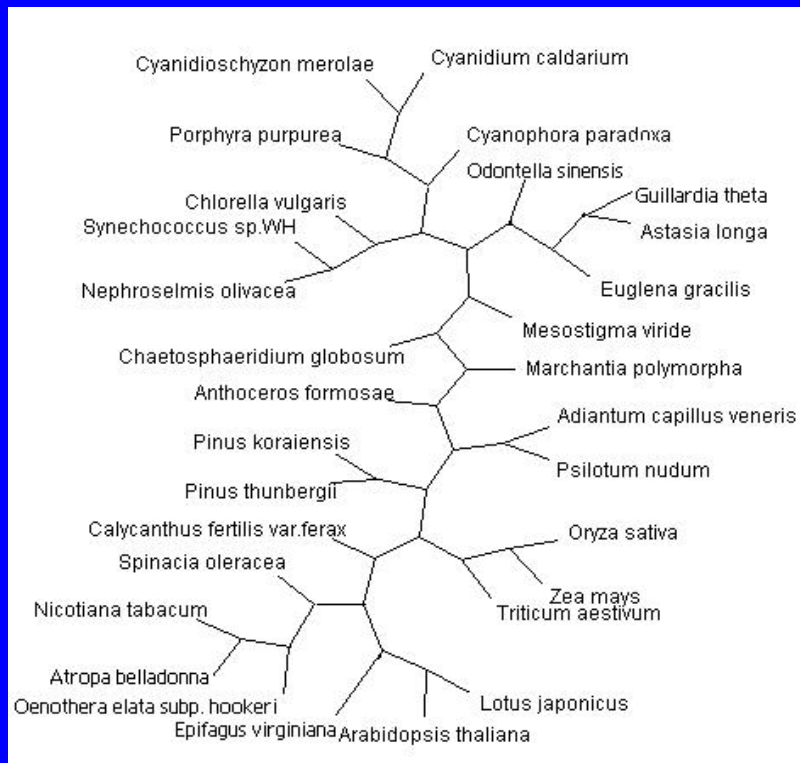
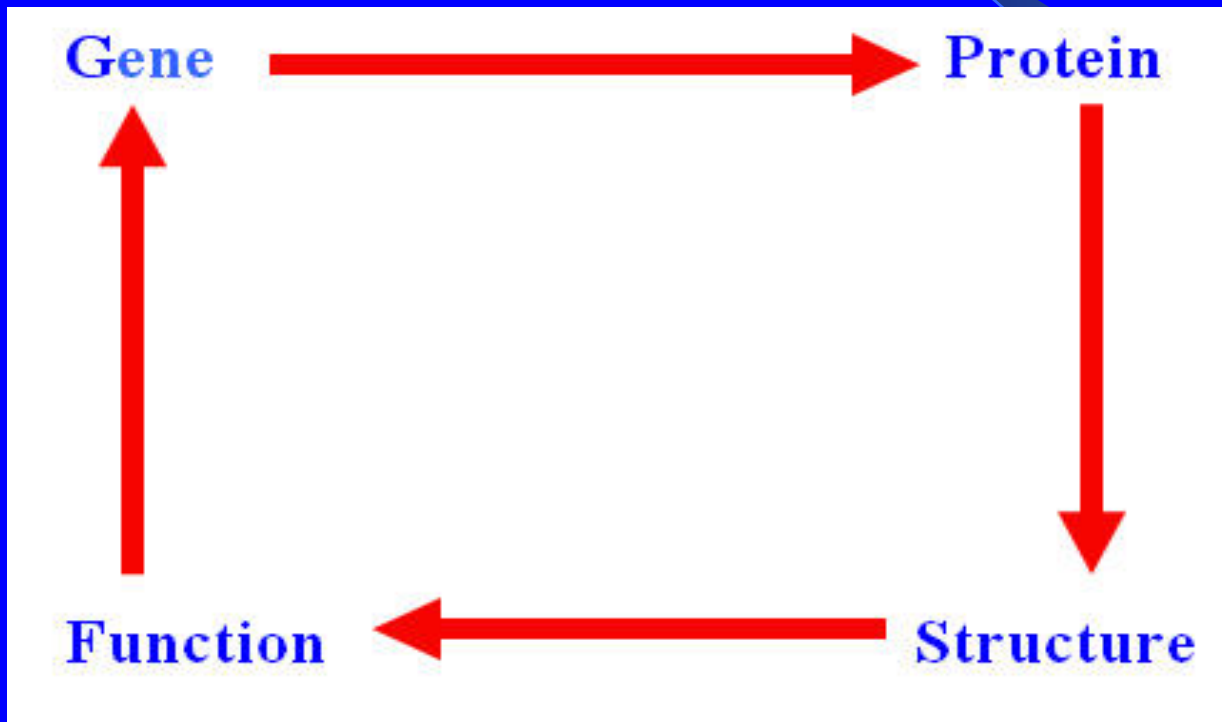


Table 2. MOP uninformative characters

Gene Name	Total # of characters	Characters deleted	# of characters used
atpA	2687	1825	862
atpB	3241	2494	747
psbA	2972	2533	439
psbC	12445	11797	648
psbD	10954	10511	443
psbE	928	800	128
psbF	1285	1209	76
rpl14	489	219	270
rpl16	1729	559	1170
rpl2	1706	785	921
rpl20	410	111	291
rpl36	230	157	73
rpoB	5036	2365	2671
rps11	608	299	309
rps14	413	177	236
rps2	1456	840	616
rps3	1237	628	609
rps4	760	266	494
rps8	1153	818	335

atpA-ATP synthase CF1 alpha chain, *atpB*-ATP synthase CF1 beta chain, *psbA*-Photosystem II Q(b) protein (D1), *psbC*- Photosystem II 44 kDa apoprotein (P6), *psbD*-Photosystem II D2 protein, *psbE*-Cytochrome b559 alpha chain, *psbF*-Cytochrome b559 beta chain, *rpl2*-50S ribosomal protein L2, *rpl14*-50S ribosomal protein L14, *rpl16*-50S ribosomal protein L16, *rpl20*-50S ribosomal protein L20, *rpl36*-50S ribosomal protein L36, *rpoB*-DNA-directed RNA polymerase beta chain, *rps2*-30S ribosomal protein S2, *rps3*-30S ribosomal protein S3, *rps4*-30S ribosomal protein S4, *rps8*-30S ribosomal protein S8, *rps11*- 30S ribosomal protein S11, *rps14*- 30S ribosomal protein S14

Basically



Annotation pitfalls

- incomplete predictions
 - missed genes or exons
- mis-predictions
 - psuedogenes
- circular predictions
 - similar to predicted...
- Definition of new functional annotations from propagated mistakes within the sequence databases

Applications

- Understanding quantitative traits
- Comparative genomics to cotton, potatoes, sorghum and pearl millet not fully sequenced
- Microarray technology for gene expression relationships as well as validate genes and gene combinations
- Introduction of new genes through chloroplasts instead of nucleus in transgenics

Conclusions

- Precise gene prediction systems can effectively combine genomic sequence comparisons (comparative genomics)
- Better methods for displaying and browsing genomic sequence now possible at whole genome level
- Visualization and interpretation of outputs