Functional Annotations

Gail Rosen

Functional Annotation

- **dictionary definition of "to annotate":**
  - "to make or furnish critical or explanatory notes or comment"
- **some of what this includes for genomics**
  - gene product names
  - functional characteristics of gene products
  - physical characteristics of gene/protein/genome
  - overall metabolic profile of the organism
- **elements of the annotation process**
  - gene finding
  - homology searches
  - functional assignment
  - ORF management
  - data availability

ORF

- Open Reading Frame

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

Sample sequence showing three different reading frames. Start codons are highlighted in purple, and stop codons are highlighted in red.

Annotation Pipeline

Generation of Open Reading Frames

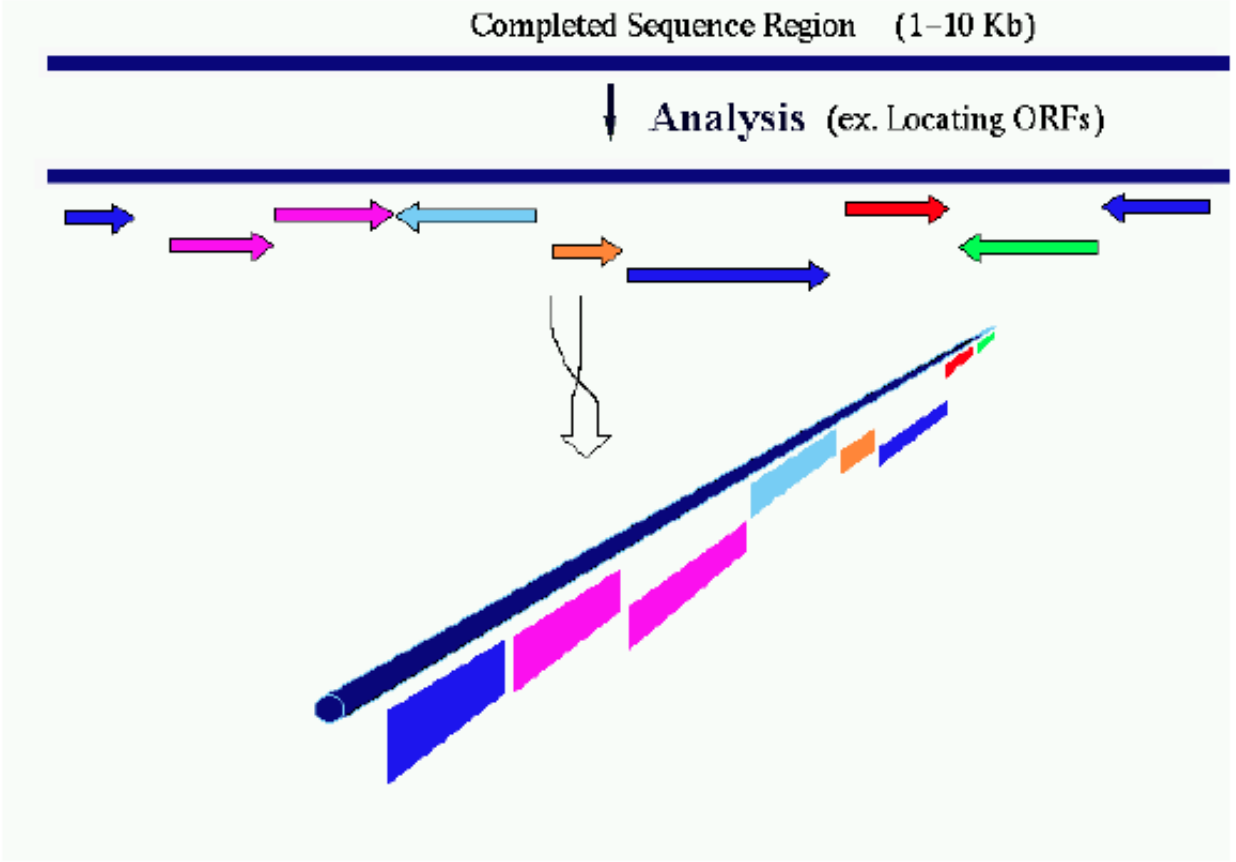
Homology Searches

Putative ID
Frameshift Detection
Ambiguity Report

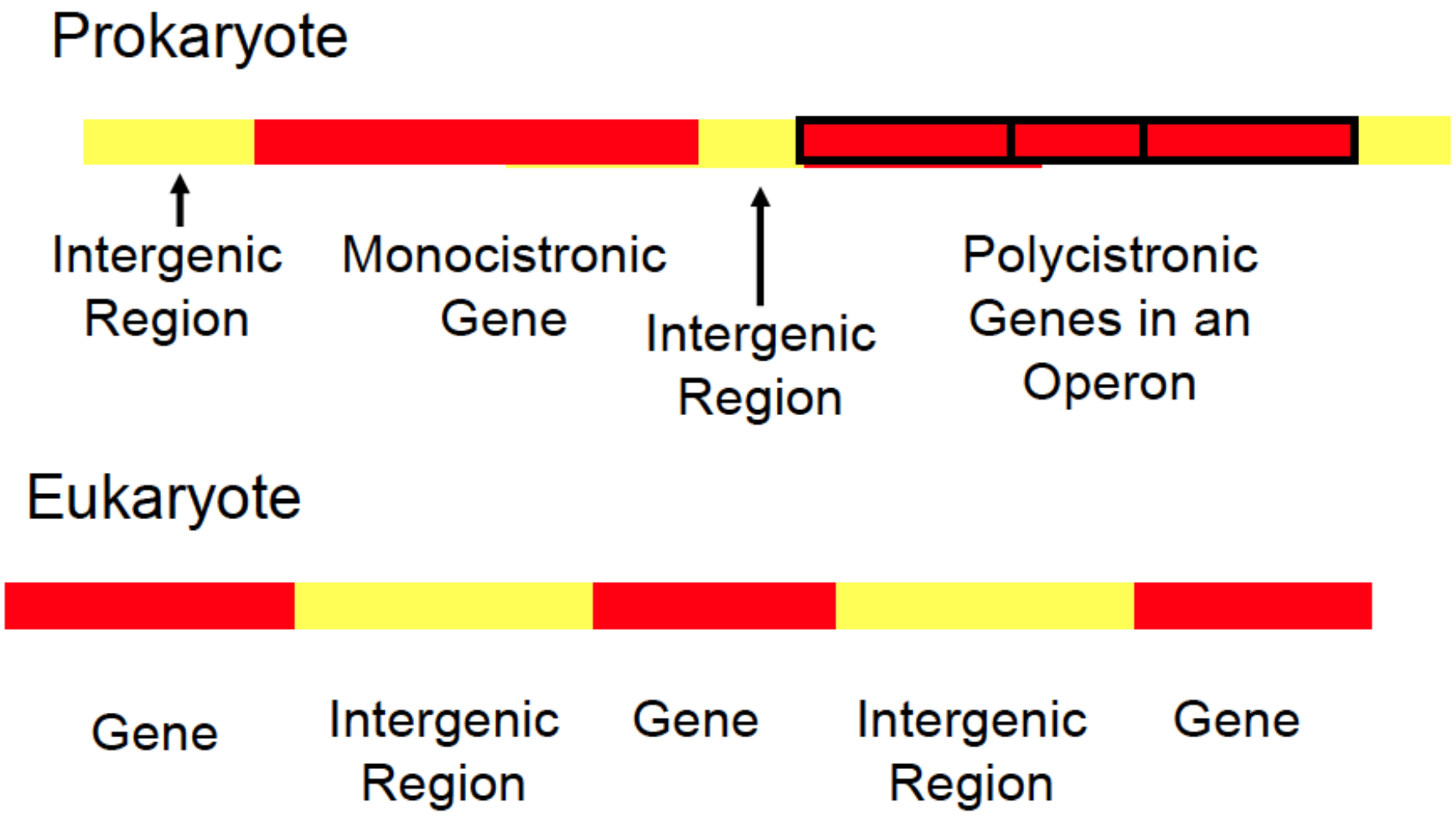
Role Assignment
Metabolic Pathways
Gene Families

DNA Motifs
Regulatory Elements
Repetitive Sequences

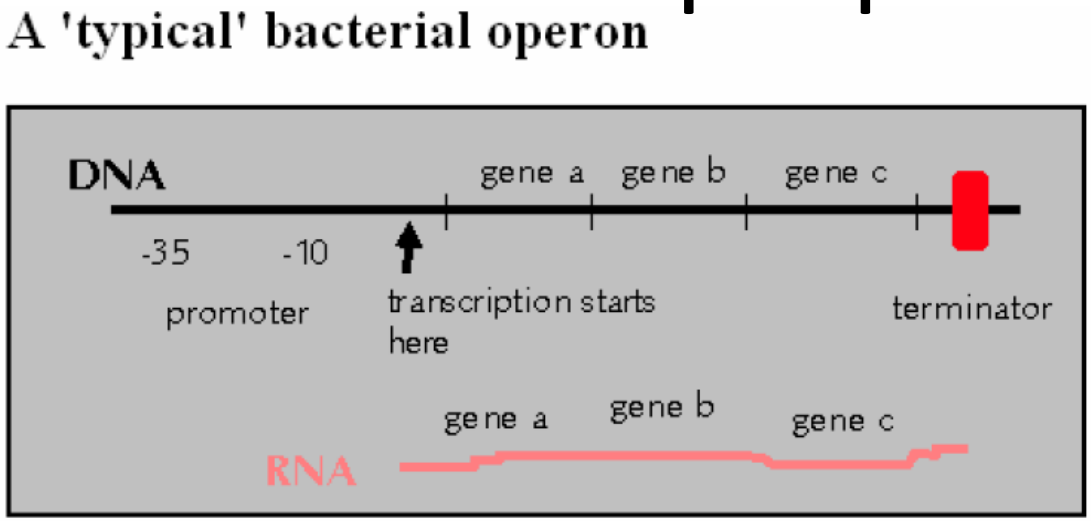
Comparative Genomics

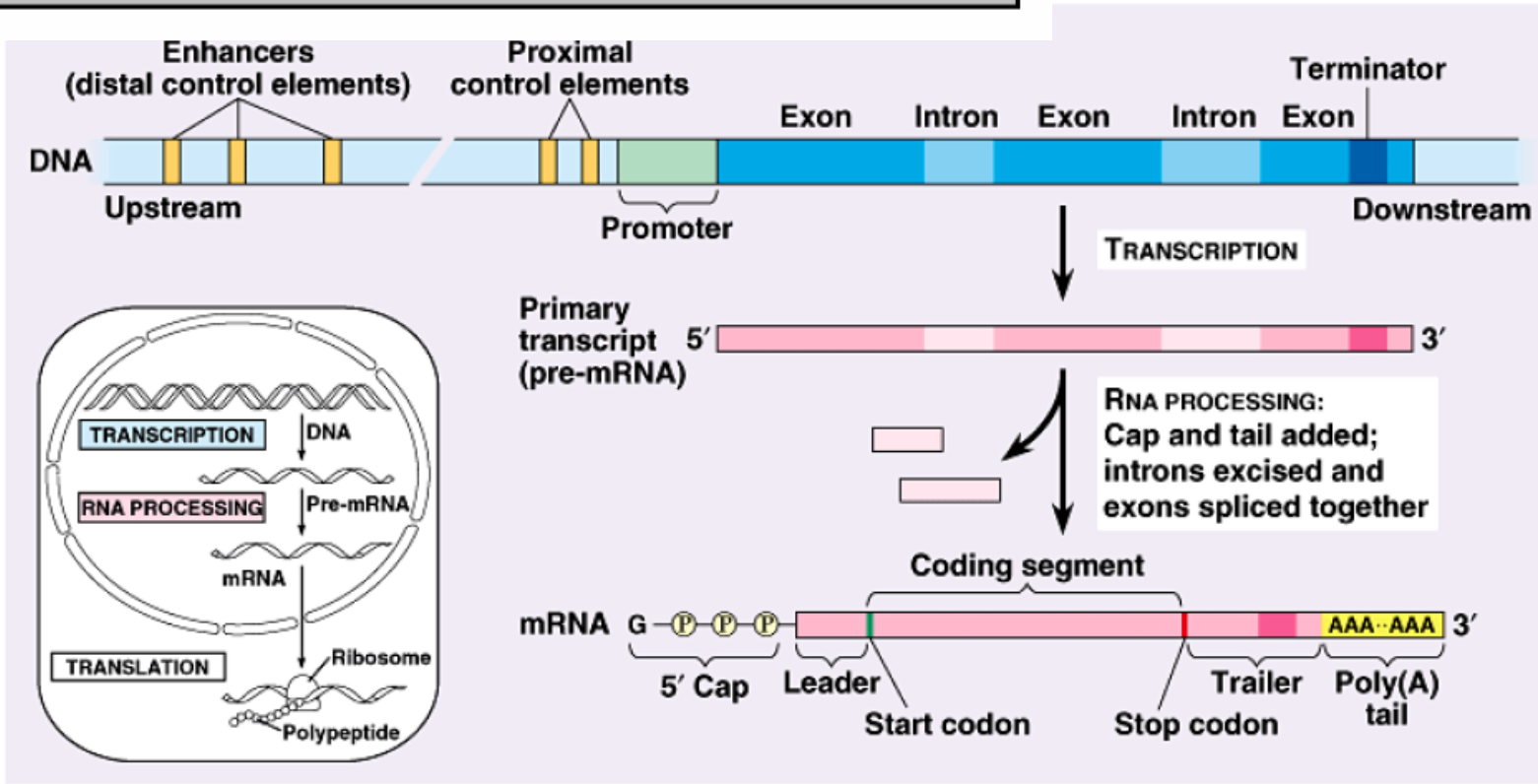


Genome Structure



Transcript “processing”





Annotating

Two main types of data used in defining gene structure:

Prediction based: algorithms designed to find genes/gene structures based on nucleotide sequence and composition

Sequence similarity (DNA and protein): alignment to mRNA sequences (ESTs) and proteins from the same species or related species; identification of domains and motifs

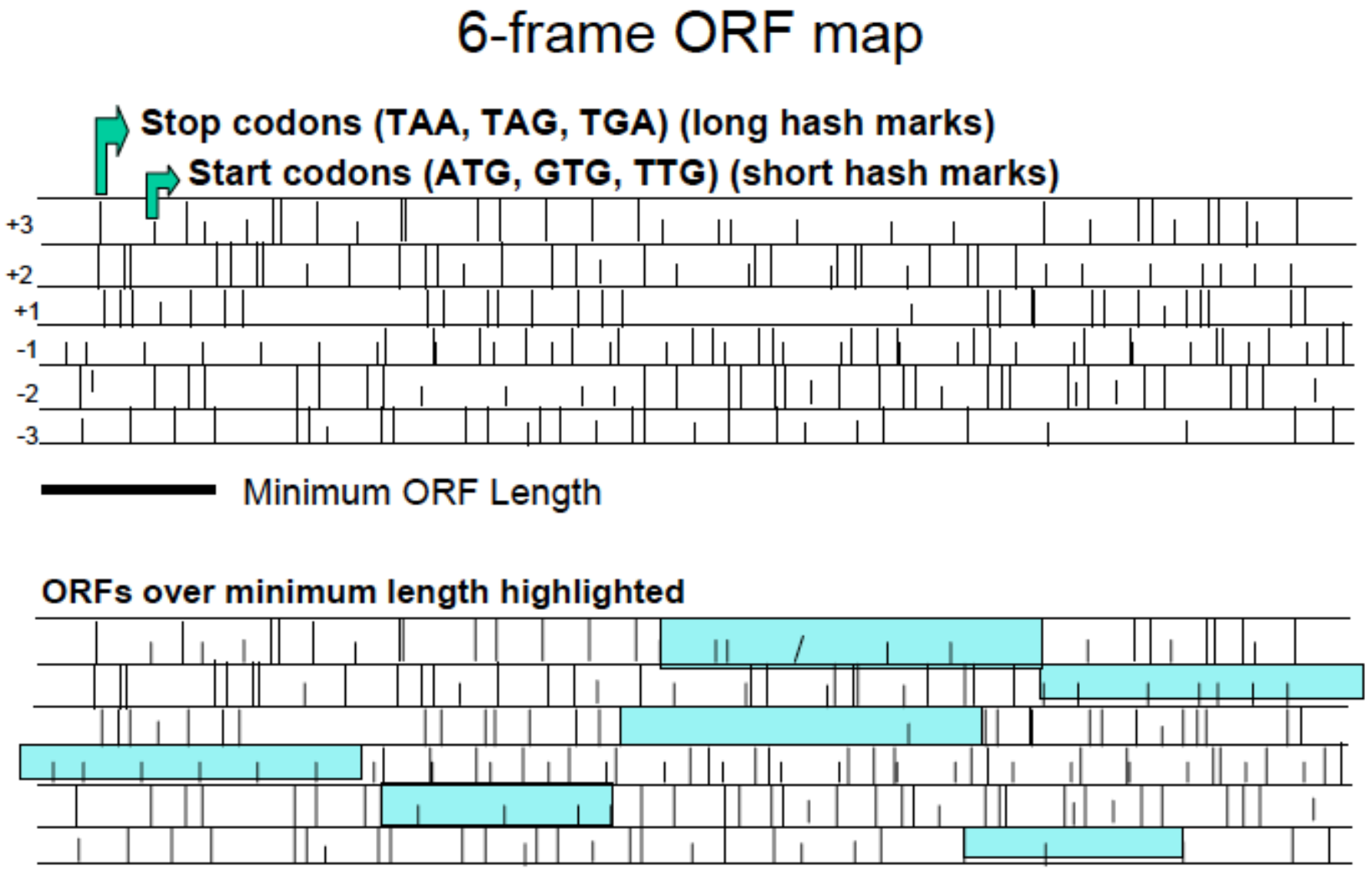
First Step: ORF Finding (traditionally whole organisms)

Running a Gene-finder is a two-part process

1) Train Gene finder for the organism you have sequenced.

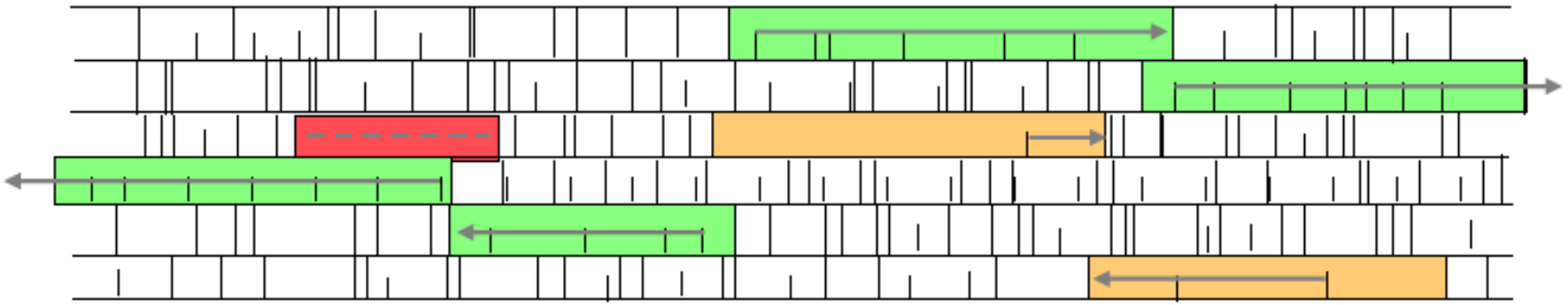
2) Run the trained Gene finder on the completed sequence.

Candidate Genes

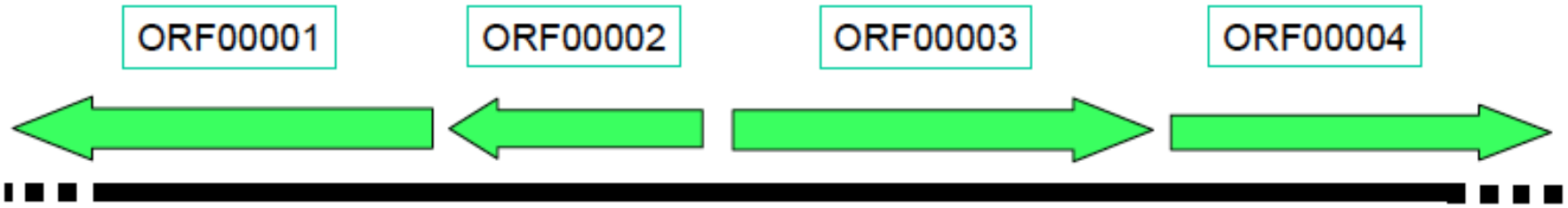


Annotate ORFs

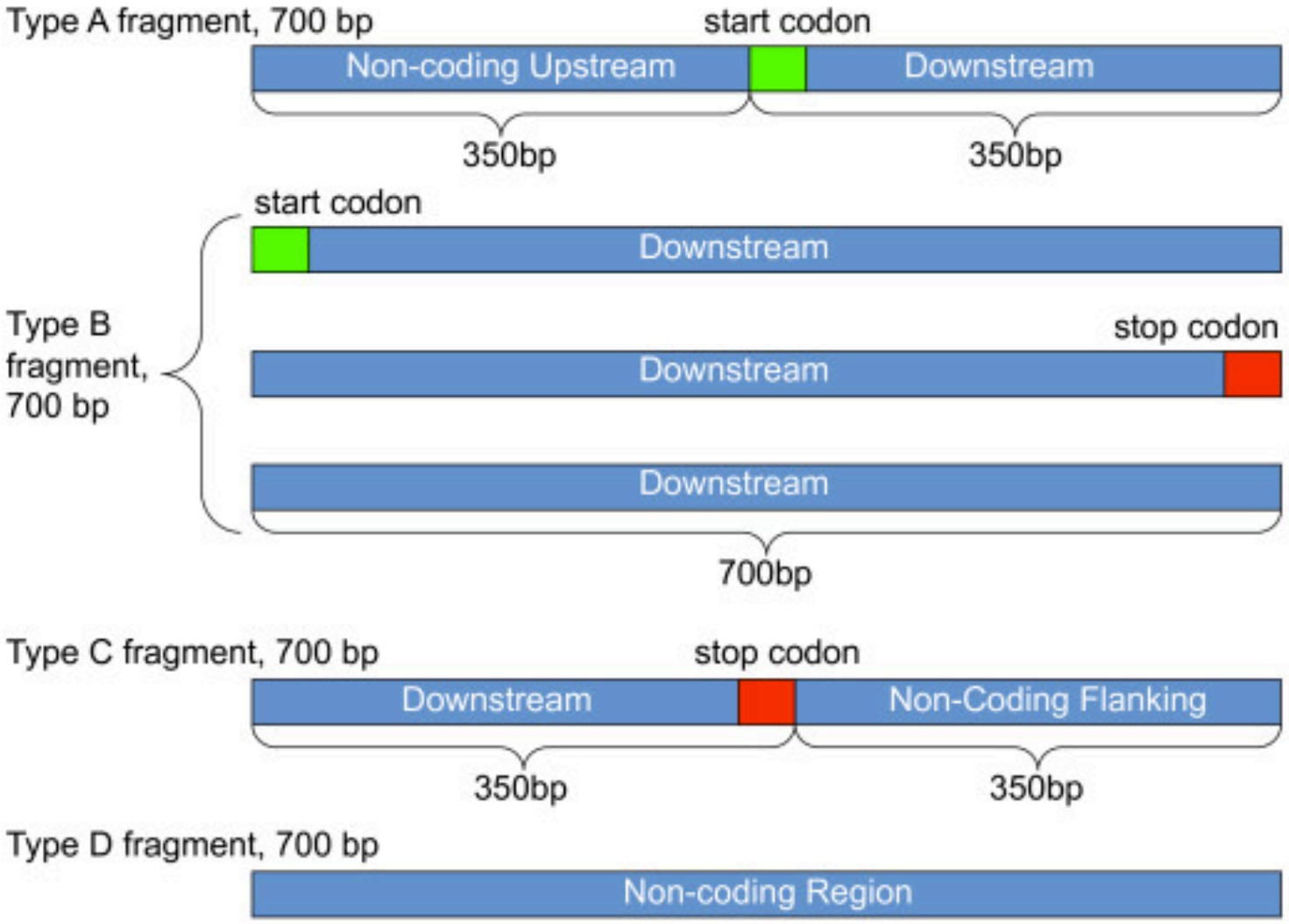
Possible translations represented by arrows, moving from start to stop, the dotted line represents an ORF with no start site.



Glimmer chooses the set of likely genes.



More Complicated for Metagenomics



Functional Assignments

Name

Descriptive common name for the protein, with as much specificity as the evidence supports; gene symbol.

Role

Describe what the protein is doing in the cell and why.

Associated information:

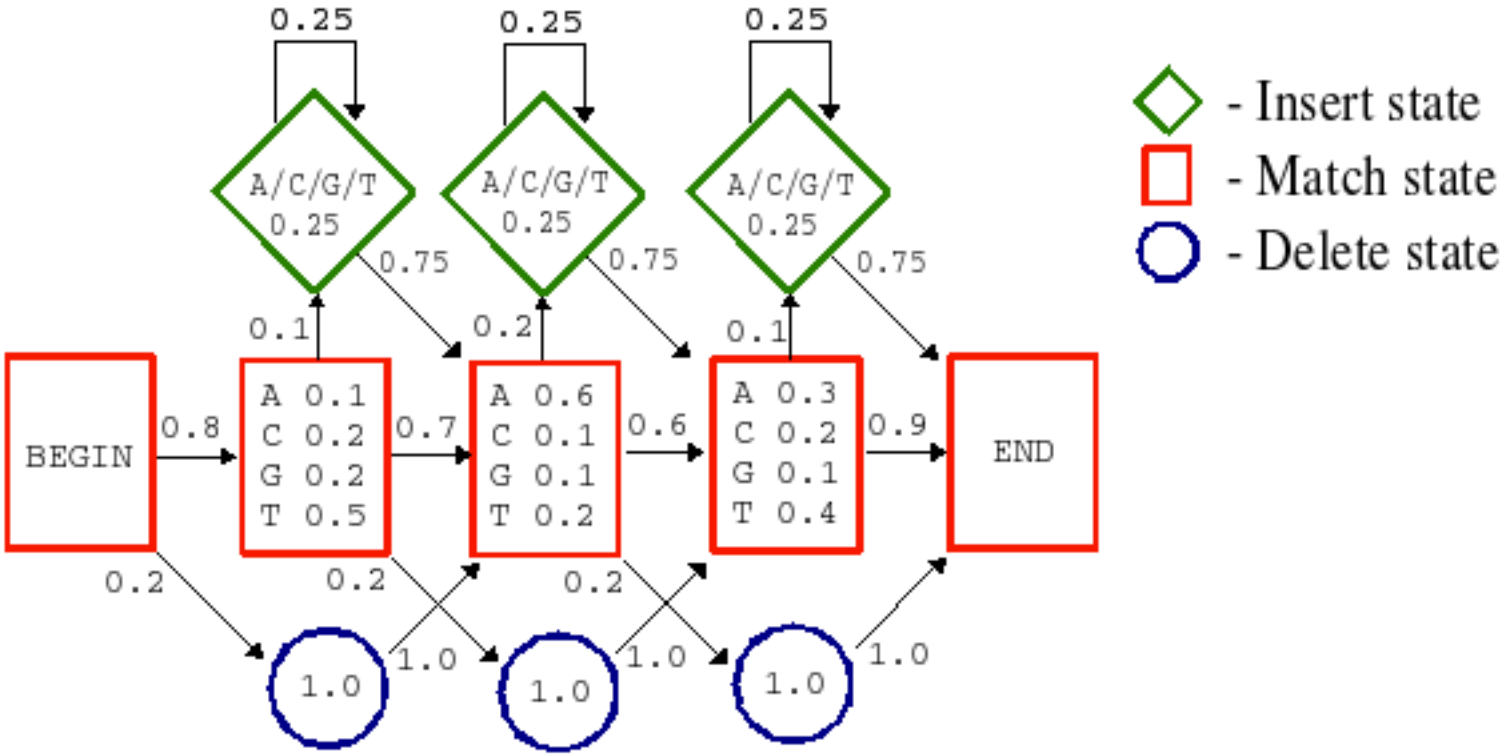
Supporting evidence:Domain and motifs
EC number if protein is an enzyme.
Paralogous family membership.

Evidence for Gene Function

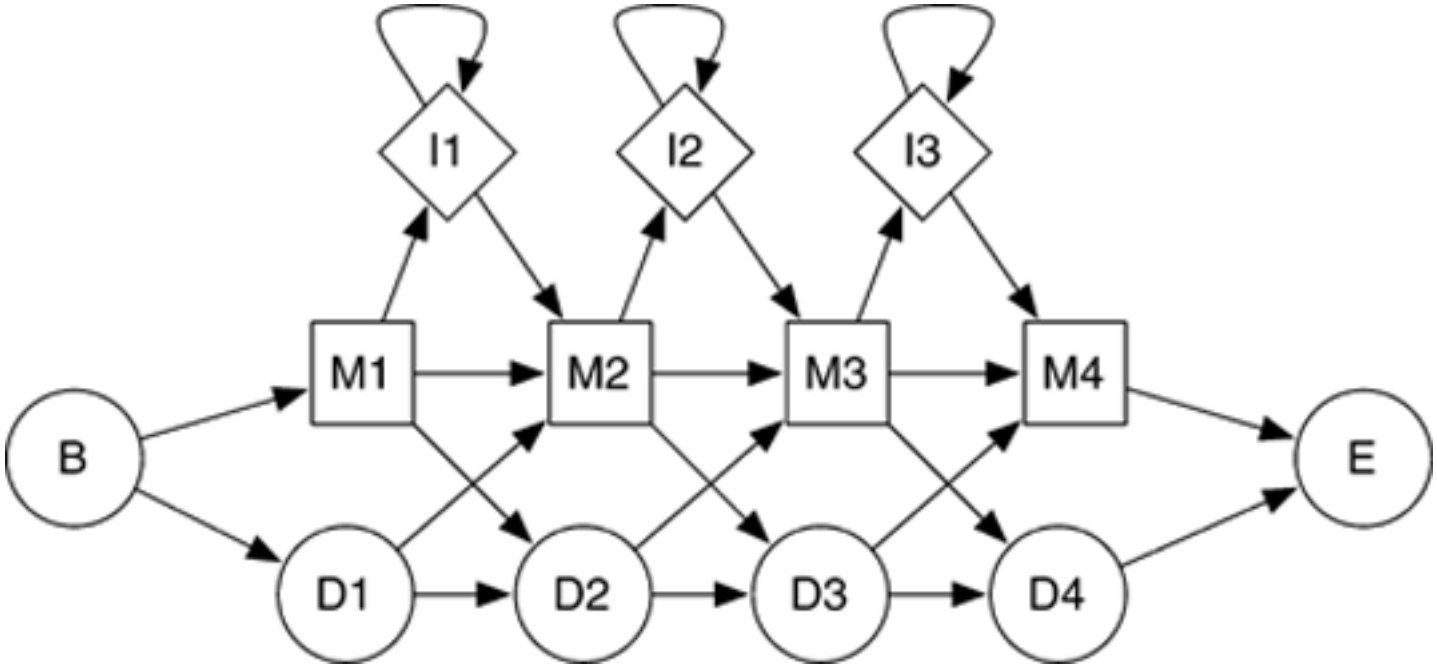
- PROSITE Motifs
  - collection of protein motifs associated with active sites, binding sites, etc.
  - help in classifying genes into functional families when HMMs for that family have not been built

- InterPro
  - Brings together HMMs (both TIGR and Pfam) Prosite motifs and other forms of motif/domain clustering
  - Results in motif “signatures” for families or functions
  - GO terms have been assigned to many of these

Markov Chains



Profile HMMs



* Viterbi: Find labels given model and sequence
* Forward/Backward Algorithm: Find Probability of label at a certain position given model and sequence

Functional annotation in practice:
searching against sequences with known functions, in various gene/protein databases:

- **gene functions** - Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- **gene orthologous groups** - COG, KOG, eggNOG, *etc.*
- **protein domains/motifs** - Pfam, FIGfam, TIGRfam, *etc.*
- **pathways/subsystems** - KEGG, MetaCyc, SEED, *etc.*

- ***gene functions*** - Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- ***gene orthologous groups*** - COG, KOG, eggNOG, *etc.*
- ***protein domains/motifs*** - Pfam, FIGfam, TIGRfam, *etc.*
- ***pathways/subsystems*** - KEGG, MetaCyc, SEED, *etc.*

Manually curated annotations for > 500,000 sequences, each with a 6-digit unique ID.

P14567 (DPO3A_SALTY) Reviewed, UniProtKB/Swiss-Prot

Names and origin

Protein names	*Recommended name:* **DNA polymerase III subunit alpha** EC=2.7.7.7
Gene names	Name: **dnaE** Synonyms: polC Ordered Locus Names: STM0231
Organism	**Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)** [Reference proteome] [HAMAP]
Taxonomic identifier	99287 [NCBI]
Taxonomic lineage	Bacteria › Proteobacteria › Gammaproteobacteria › Enterobacteriales › Enterobacteriaceae › Salmonella ›

Protein attributes

Sequence length	1160 AA.
Sequence status	Complete.
Protein existence	Inferred from homology

General annotation (Comments)

Function	DNA polymerase III is a complex, multichain enzyme responsible for most of the replicative synthesis in bacteria. This DNA polymerase also exhibits 3' to 5' exonuclease activity. The alpha chain is the DNA polymerase.
Catalytic activity	Deoxynucleoside triphosphate + DNA(n) = diphosphate + DNA(n+1).
Subunit structure	The DNA polymerase holoenzyme is a complex that contains 10 different types of subunits. These subunits are organized into 3 functionally essential subassemblies: the pol III core, the beta sliding clamp processivity factor and the clamp-loading complex. The pol III core (subunits alpha,epsilon and theta) contains the polymerase and the 3'-5' exonuclease proofreading activities. The polymerase is tethered to the template via the sliding clamp processivity factor. The clamp-loading complex assembles the beta processivity factor onto the primer template and plays a central role in the organization and communication at the replication fork. This complex contains delta, delta', psi and chi, and copies of either or both of two different DnaX proteins, gamma and tau. The composition of the holoenzyme is, therefore: (alpha,epsilon,theta)[2]-(gamma/tau)[3]-delta,delta', psi,chi-beta[4].
Subcellular location	Cytoplasm (By similarity).
Sequence similarities	Belongs to the DNA polymerase type-C family. DnaE subfamily.

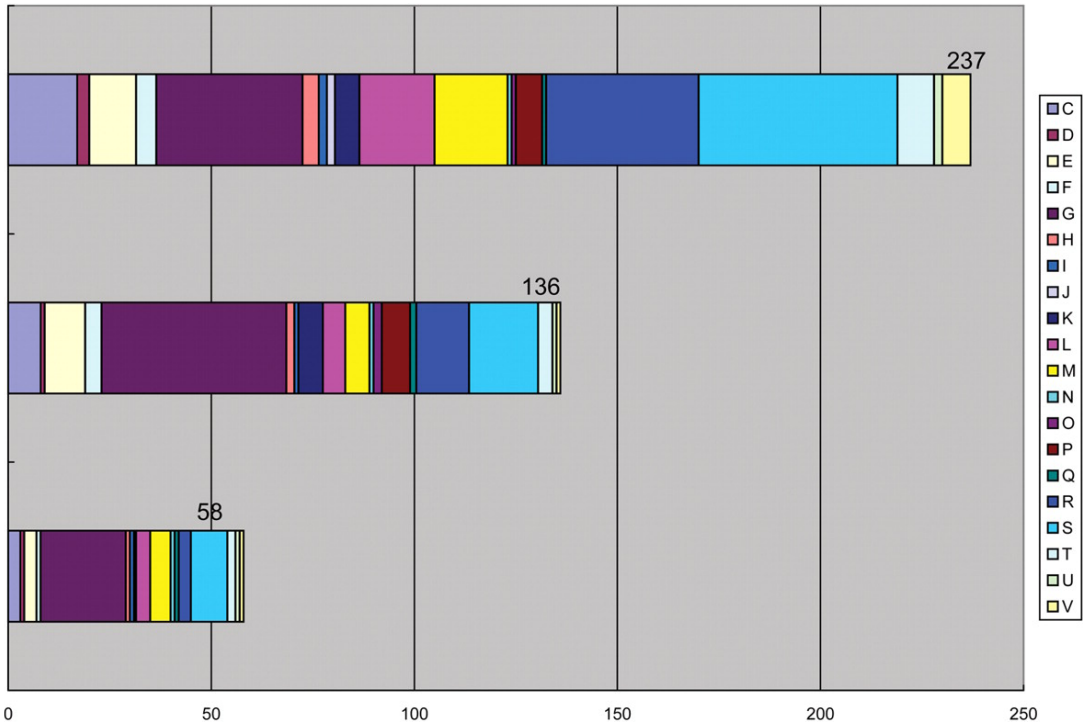
- ***gene functions*** **-** Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- ***gene orthologous groups*** **-** COG, KOG, eggNOG, *etc.*
- ***protein domains/motifs*** **-** Pfam, FIGfam, TIGRfam, *etc.*
- ***pathways/subsystems*** **-** KEGG, MetaCyc, SEED, *etc.*

Clusters of Orthologous Groups

Genes are grouped into 23 categories

A	RNA processing and modification
B	Chromatin Structure and dynamics
C	Energy production and conversion
D	Cell cycle control and mitosis
E	Amino Acid metabolis and transport
F	Nucleotide metabolism and transport
G	Carbohydrate metabolism and transport
H	Coenzyme metabolis
I	Lipid metabolism
J	Tranlsation
K	Transcription
L	Replication and repair
M	Cell wall/membrane/envelop biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, chaperone functions
P	Inorganic ion transport and metabolism
Q	Secondary Structure
T	Signal Transduction
U	Intracellular trafficing and secretion
Y	Nuclear structure
Z	Cytoskeleton
R	General Functional Prediction only
S	Function Unknown

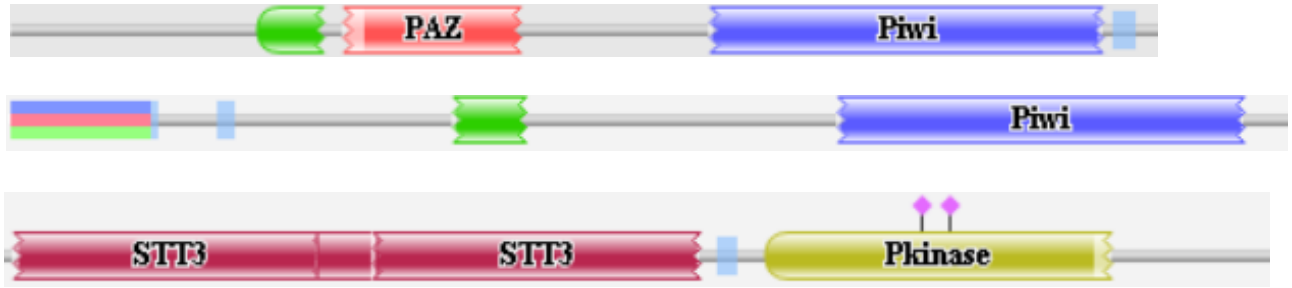
COGs for enriched genes in "adult/children", "infant" and "both"



Kurokawa et al., 2007

- ***gene functions*** **-** Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- ***gene orthologous groups*** **-** COG, KOG, eggNOG, *etc.*
- ***protein domains/motifs*** **-** Pfam, FIGfam, TIGRfam, *etc.*
- ***pathways/subsystems*** **-** KEGG, MetaCyc, SEED, *etc.*

Protein domains on sequences that are 25 up to 500 amino acids in length.



PAZ domain

Provide feedback

This domain is named PAZ after the proteins Piwi Argonaut and Zwille. This domain is found in two families of proteins that are involved in post-transcriptional gene silencing. These are the Piwi family and the Dicer family, that includes the Carpel factory protein. The function of the domains is unknown but has been suggested to mediate complex formation between proteins of the Piwi and Dicer families by hetero-dimerisation. The three-dimensional structure of this domain has been solved [2-4]. The PAZ domain is composed of two subdomains. One subdomain is similar to the OB fold, albeit with a different topology. The OB-fold is well known as a single-stranded nucleic acid binding fold. The second subdomain is composed of a beta-hairpin followed by an alpha-helix. The PAZ domains shows low-affinity nucleic acid binding and appears to interact with the 3' ends of single-stranded regions of RNA in the cleft between the two subdomains. PAZ can bind the characteristic two-base 3' overhangs of siRNAs, indicating that although PAZ may not be a primary nucleic acid binding site in Dicer or RISC, it may contribute to the specific and productive incorporation of siRNAs and miRNAs into the RNAi pathway.

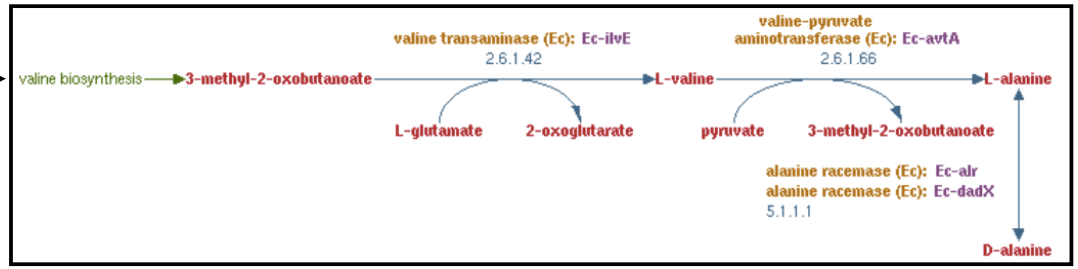
Literature references

1. Cerutti L, Mian N, Bateman A; , Trends Biochem Sci 2000;25:481-482.: Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. PUBMED:11050429 EPMC:11050429
2. Song JJ, Liu J, Tolia NH, Schneiderman J, Smith SK, Martienssen RA, Hannon GJ, Joshua-Tor L; , Nat Struct Biol 2003;10:1026-1032.: The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. PUBMED:14625589 EPMC:14625589

- ***gene functions*** **-** Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- ***gene orthologous groups*** **-** COG, KOG, eggNOG, *etc.*
- ***protein domains/motifs*** **-** Pfam, FIGfam, TIGRfam, *etc.*
- ***pathways/subsystems*** **-** KEGG, MetaCyc, SEED, *etc.*

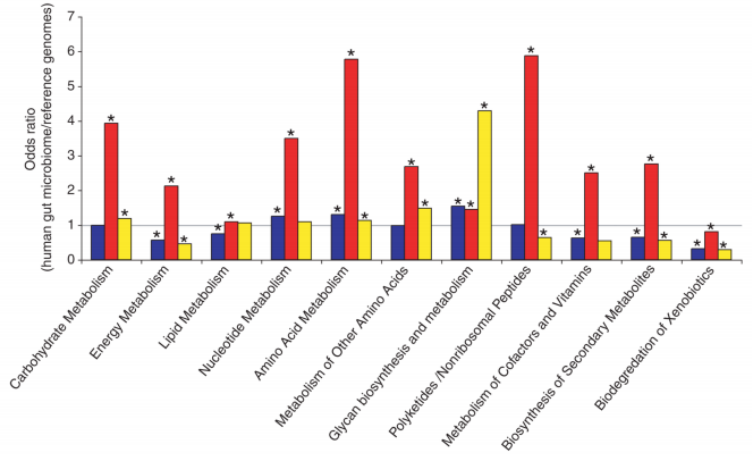
Experimentally elucidated metabolic pathways

- Pathways
  - Activation/Inactivation/Interconversion (34 instances)
  - Biosynthesis (1401 instances)
    - Amines and Polyamines Biosynthesis (38 instances)
    - Amino Acids Biosynthesis (116 instances)
      - Individual Amino Acids Biosynthesis (103 instances)
        - Alanine Biosynthesis (4 instances)
          - alanine biosynthesis I
          - alanine biosynthesis II
          - alanine biosynthesis III
          - superpathway of alanine biosynthesis
        - Arginine Biosynthesis (5 instances)
        - Asparagine Biosynthesis (5 instances)
        - Aspartate Biosynthesis (2 instances)
  - Degradation/Utilization/Assimilation (896 instances)
  - Detoxification (40 instances)
  - Generation of Precursor Metabolites and Energy (192 instances)
  - Metabolic Clusters (36 instances)
  - Signal transduction pathways (0 instances)
  - Superpathways (351 instances)
  - Transport (0 instances)

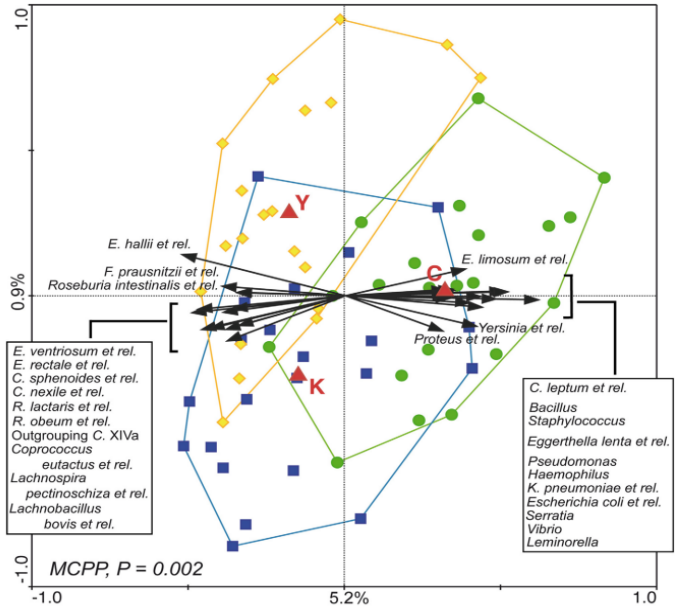


- ***gene functions*** **-** Gene Ontology, NCBI RefSeq, UniProtKB, *etc.*
- ***gene orthologous groups*** **-** COG, KOG, eggNOG, *etc.*
- ***protein domains/motifs*** **-** Pfam, FIGfam, TIGRfam, *etc.*
- ***pathways/subsystems*** **-** KEGG, MetaCyc, SEED, *etc.*

	function1	function2	function3	…
metagenome1	1	0	30	
metagenome2	0	0	55	
metagenome3	1	49	5	
metagenome4	1	1	19	
…				

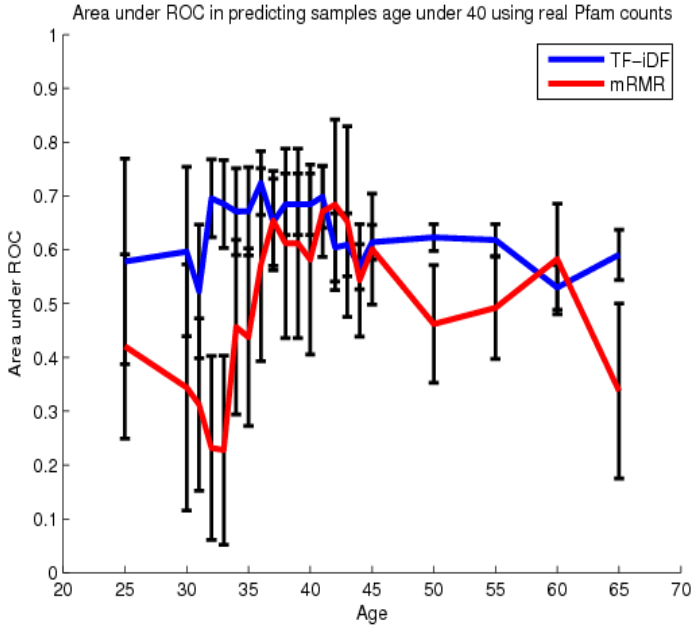


Gill, S. R., M. Pop, et al. (2006)



Biagi, E., L. Nylund, et al. (2010)

Pfam



Lan, et al. (2010)