The Toot Suite Project: Predicting and classifying membrane and transport proteins to study host-microbiome interactions

Gregory Butler

Department of Computer Science & Software Engineering
Centre for Structural and Functional Genomics
Concordia University, Montréal, Canada

October 2019 — Malaysia

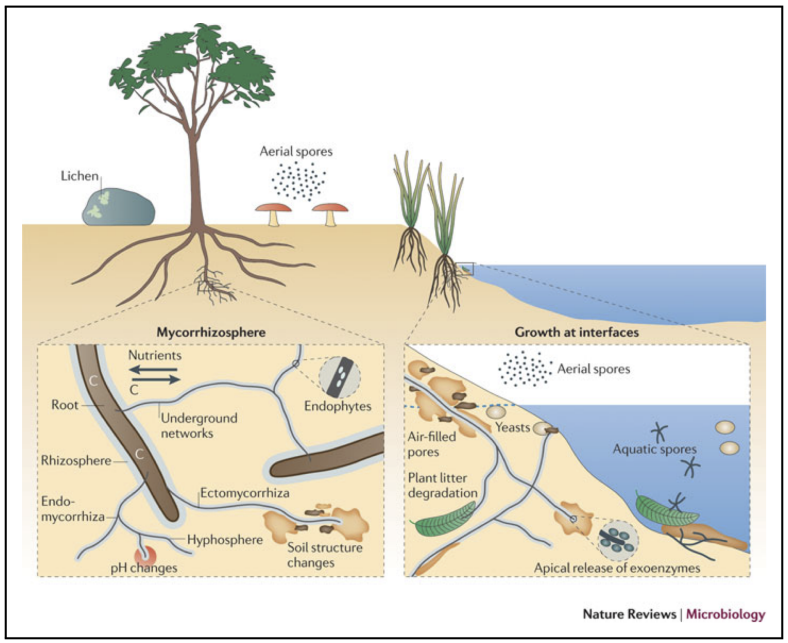
Outline

Outline

Context

- Context — Bioinformatics
- Context — Network Reconstruction
- Context — Host-Microbiome Interactions
- Context — Feeding the World

Fungi



Symbiosis

- plant roots
- lichen
- *"noble rot"*
- microbiome

Pathogens

- Plant *blight, smut, mould*
  red pine beetle
- Human *aspergillosis*, *C. albicans*
- Bacteria, insects, frogs, animal

Food

- yeast
- edible mushrooms

Degradation

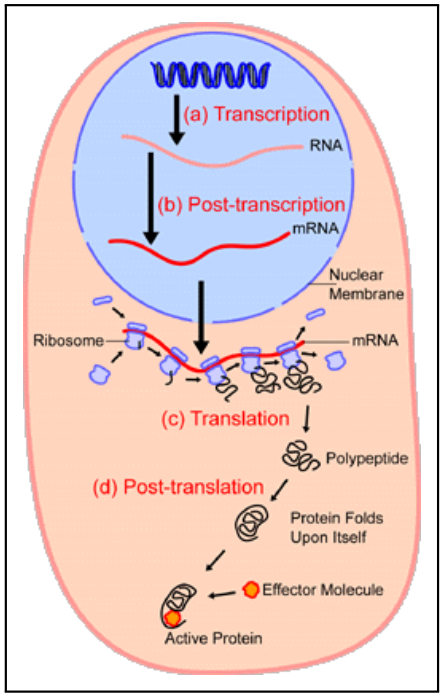
- plant litter
- polyphenols

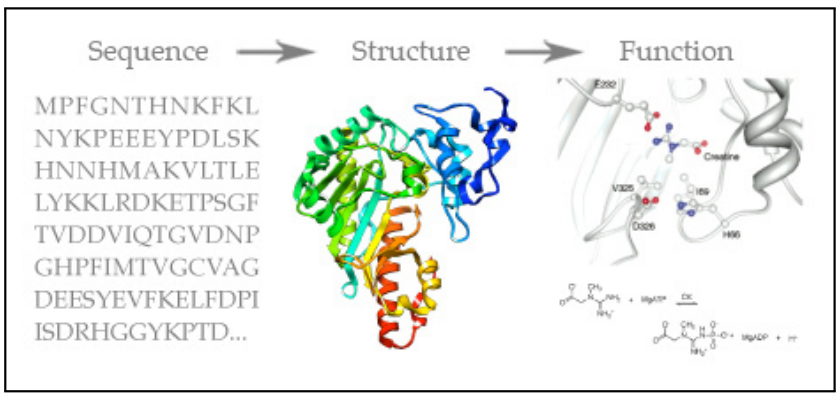
Microbiomes

- cattle rumen, elk, deer, muskoxen, etc
- termite gut

Bioinformatics in a Nutshell — Algorithmics

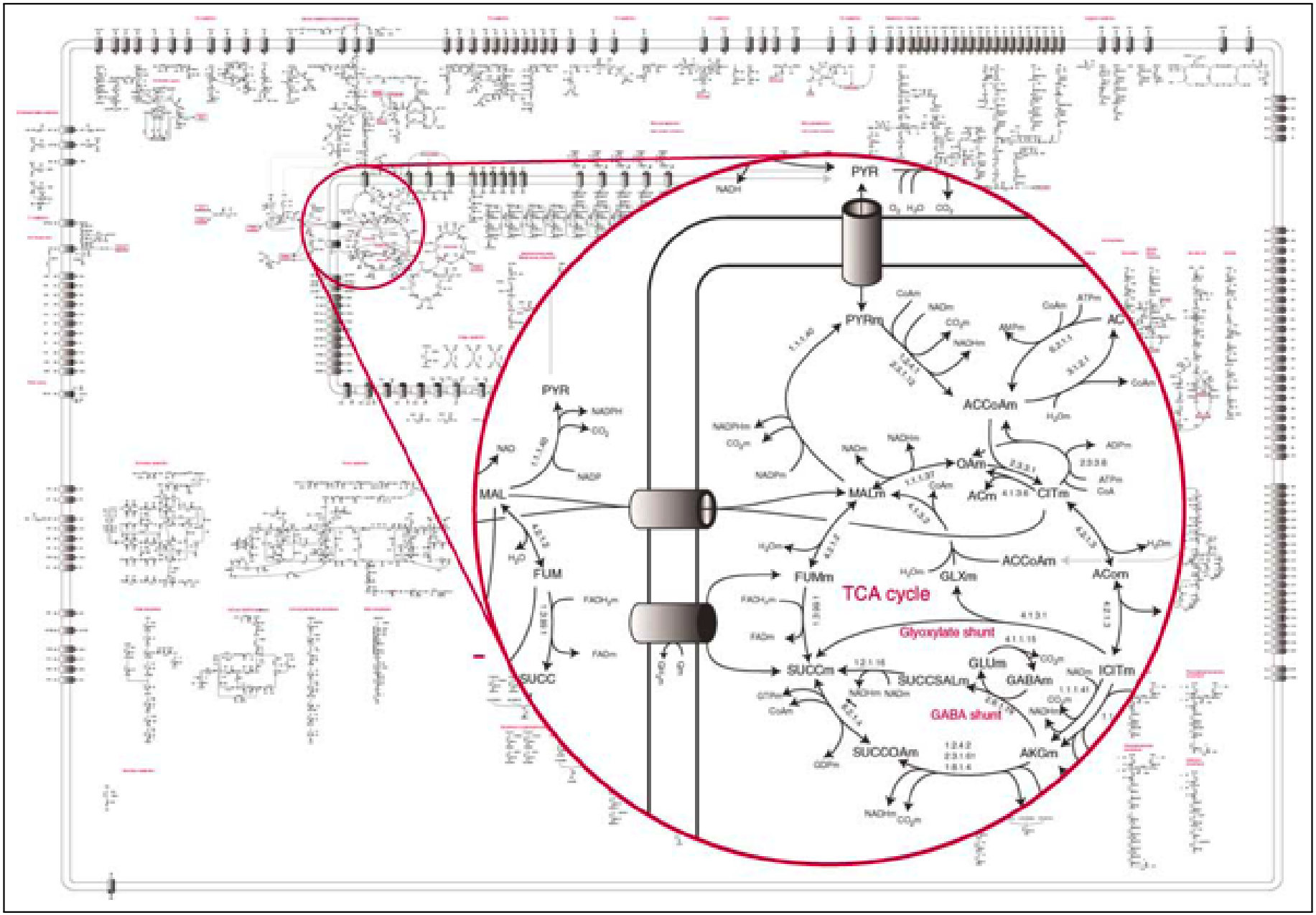
Central Dogma of Genomics



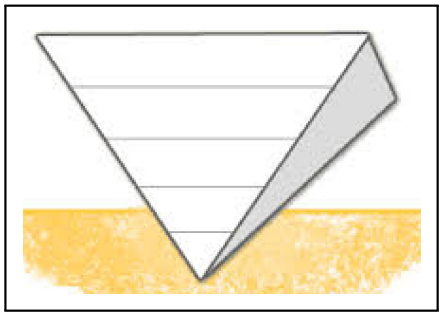


- Assembly
  from DNA reads to chromosomes
  from RNA reads to transcripts
- Structural Annotation: find genes
- Functional Annotation
  describe roles of genes (GO)
  - biological process (BP)
  - molecular function (MF)
  - cellular component (CC)
- Analysis of "-omics" expression data
  - transcriptomics
  - proteomics
  - metabolomics
- Systems Biology *holistic* perspective
  - metabolism
  - transport
  - regulation
  - signaling

Example of Metabolism and Transport



Foundation Data: Basis of Annotation



Sequenced org.: $< 1\%$ genes annotated
Well-studied org.: $\sim 10\%$ genes annotated
Model organisms: $\sim 40\%$ genes annotated

Annotation is ... Propogation of annotation by
- Annotation transfer by homology (ATH)
- Guilt-by-association (GBA)

Must track provenance

and ... Catching systematic errors by
- rule-based post-processing

Errors *often* due to phylogenetic differences, or confusion of orthologs, paralogs, xenologs

Issues with Predicting Transport

State of the art for transporter prediction is poor

Similarity works as well as any other technique

but limited to known transporters

De novo transporter prediction is poor

coverage of known transporters is low
often vastly overpredicts

Need improvements

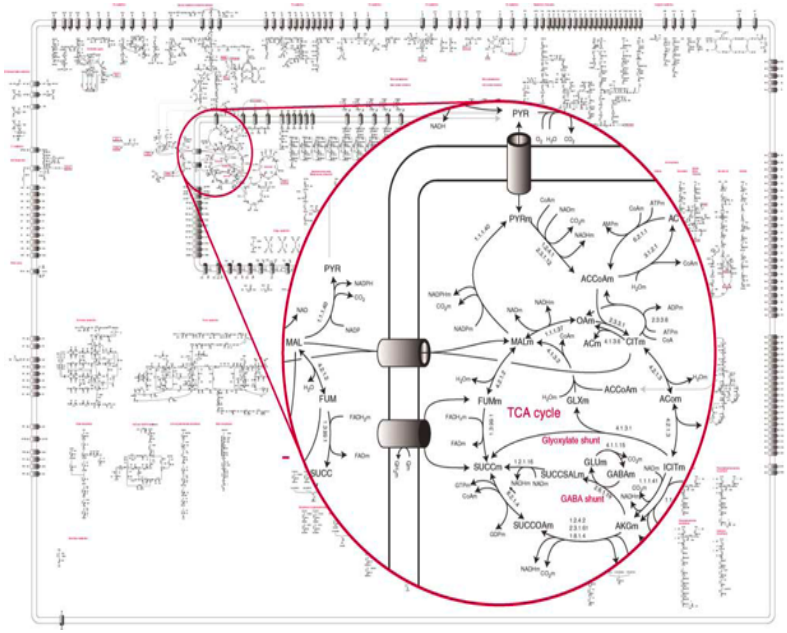
in examples of characterized transporters
in prediction of transmembrane segments (TMS)
in prediction of localization
in harmonizing classification schemes

Need to predict the specific substrate that is transported

Previous Work on Transport Prediction

Solution	Organism	Size	Substrates	Features	Classifier	Performance*
Schaadt *et al.* [70]	Specific (Arabidopsis thaliana)	61	amino acid, oligopeptides, phosphate and hexose	AAC, PAAC, PseAAC, MSA-AAC	Euclidean distance	Accuracy of 90%
Chen *et al.* [71]	General	651	electron, protein/ mRNA, ion and others	AAC, PAAC, AAindex, PSSM	Neural network	Accuracy of about 80%
Schaadt *et al.* [73]	Specific (Arabidopsis thaliana)	61	amino acid, oligopeptides, phosphate and hexose	AAC with separating TM-segments	Euclidean distance	Accuracy of 80%
Barghash *et al.* [74]	Specific (Escherichia coli, Saccharomyces cerevisiae, Arabidopsis thaliana)	246	amino acids, metal ions, phosphates and sugars	BLAST, HMMER, MEME	N/A	F-measure around 40-75%
Mishra *et al.* [77]	General	780	amino acid, anion, cation, electron, protein/ mRNA, sugar and others	AAC, PAAC, PseAAC, AAindex, PSSM	SVM	Overall MCC of 0.41 and accuracy of 78%

Manual Network Model of Andersen (2008)



Aspergillus niger CBS 513.88
14,156 ORFS

986 metabolic reactions
871 GPR associations
131 (3.14%) holes

205 transport reactions
202 (98.54%) holes

extracellular to cytosol:
151 transport reactions
cytosol to mitochondrion:
54 transport reactions

MR Andersen *et al*, Metabolic model integration of the bibliome, genome, metabolome and reactome of *Aspergillus niger*. *Molecular Systems Biology*, 4(1), 2008.

Results from Pathway Tools on Case Study

Metabolic Reactions

332 pathways, 1868 metabolic reactions, 1580 GPR associations
335 (31%) gaps

Transport Reactions Predicted

(1) $NADP^+ + NADH + H^+_{[out]} \longleftrightarrow NAD^+ + NADPH + H^+_{[in]}$
(2) $UDP{-}\alpha{-}D{-}glucose + glucosyl{-}glycogenin_{[in]} \longrightarrow$
$1,4\text{-}\alpha{-}D{-}glucosylglycogenin + UDP + H^+_{[out]}$
(3) $phospolipid_{[in]} + ATP + H_2O \longleftrightarrow phospolipid_{[out]} + ADP + phosphate + H^+$
(4) $Cu^{2+}_{[in]} + ATP + H_2O \longrightarrow Cu^{2+}_{[out]} + ADP + phosphate + H^+$
(5) $ATP + H^+_{[in]} + H_2O \longleftrightarrow ADP + phosphate + H^+_{[out]}$
(6) $Ca^{2+}_{[out]} + ATP + H_2O \longleftrightarrow Ca^{2+}_{[in]} + ADP + phosphate + H^+$
(7) $ATP + 3\,H^+_{[in]} + H_2O \longleftrightarrow ADP + phosphate + 4\,H^+_{[out]}$
(8) $oligopeptide_{[out]} + ATP + H_2O \longleftrightarrow oligopeptide_{[in]} + ADP + phosphate$
(9) $xenobiotic_{[in]} + ATP + H_2O \longleftrightarrow xenobiotic_{[out]} + ADP + phosphate$
(10) $4\,H^+_{[in]} \longrightarrow 4\,H^+_{[out]}$

Toot Suite — Motivation

Help understand host-microbiome interaction

by predicting transporter proteins
and their substrates

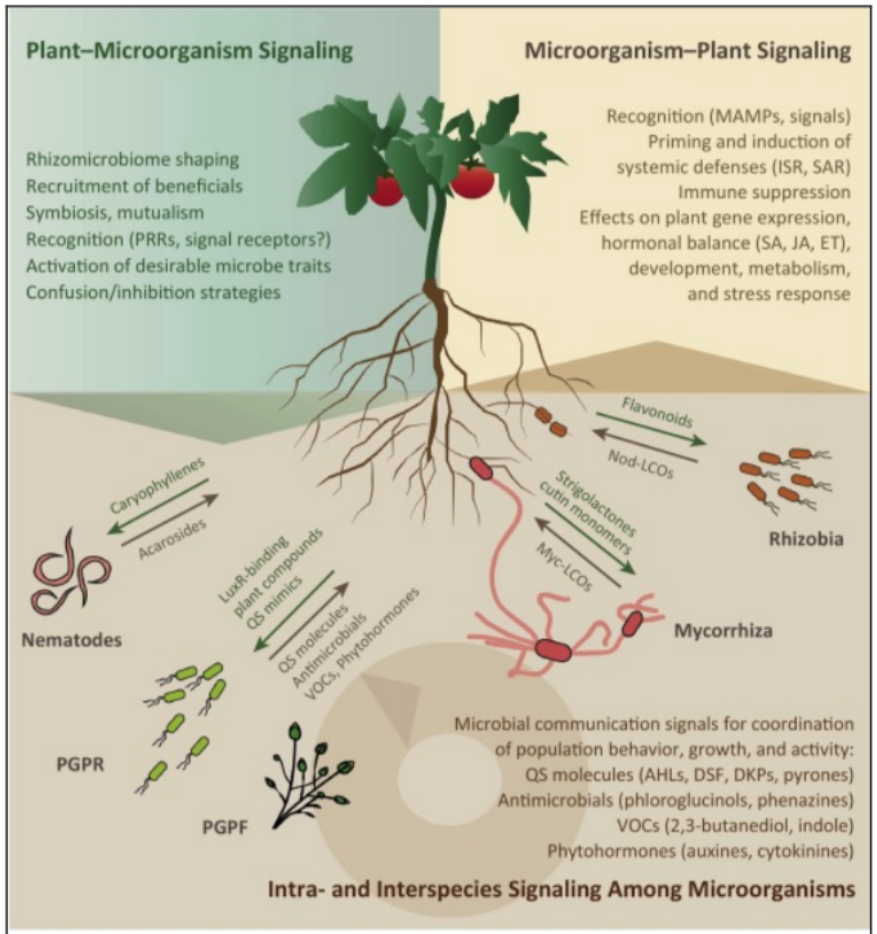


Figure 2: Plant Rhizosphere [Venturi'16]

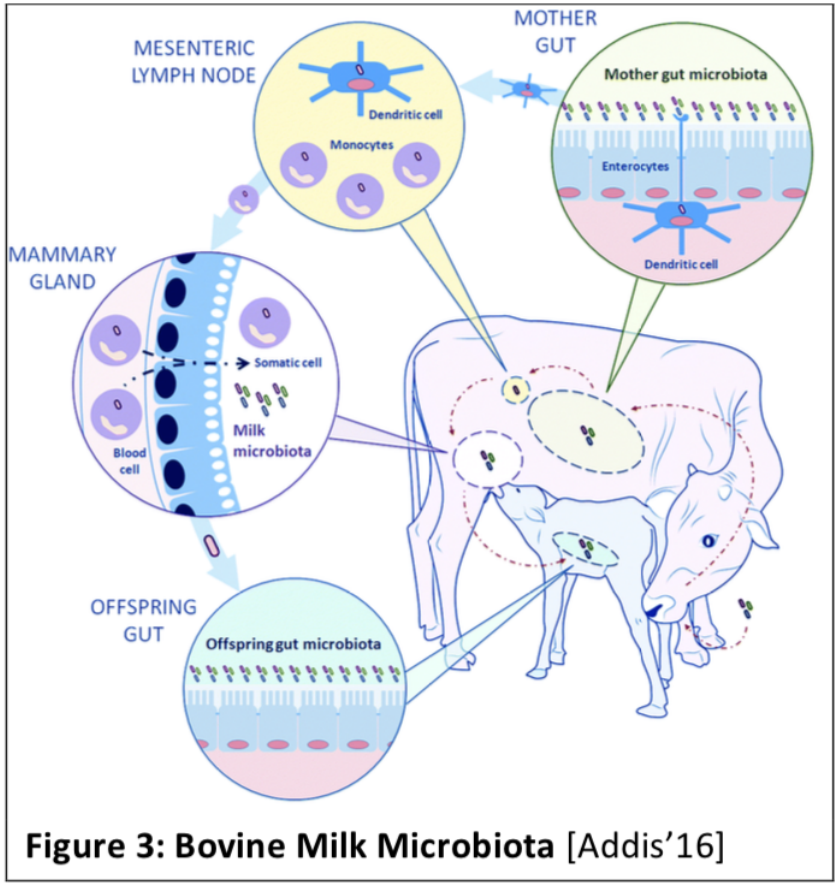


Figure 3: Bovine Milk Microbiota [Addis'16]

Outline

TooT Suite Project

The Toot Suite Project

Genome Canada BCB 2017 Competition

TooT Suite: Predication and classification of membrane transport proteins, Gregory Butler and Tristan Glatard, 2018–2021

Bioinformatics and Machine Learning

Develop predictors for transporter proteins and membrane proteins

Open Science

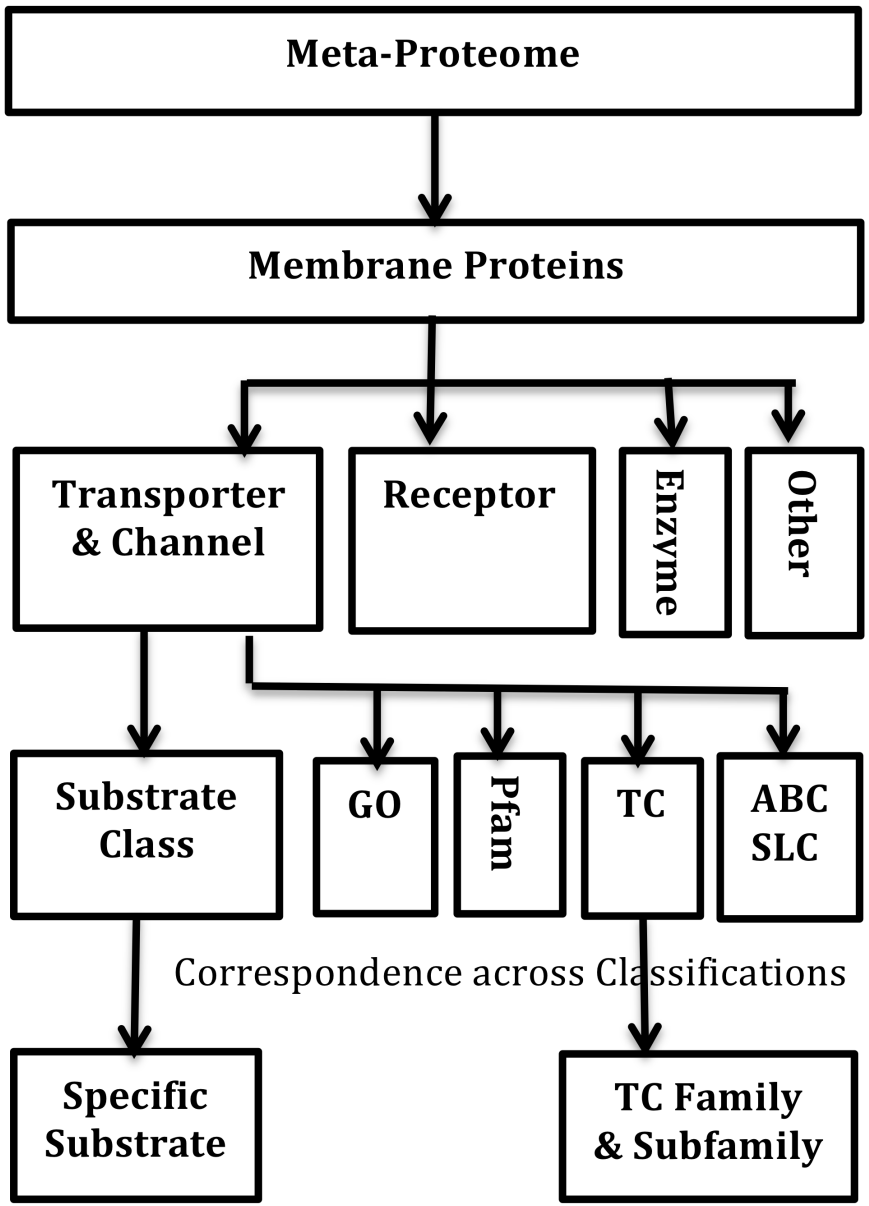
tools — open source
platform for experiments — Boutiques + bfx tools + ML tools
reproducible experiments

Scale to microbiomes

Motivation

Improve agricultural productivity
provide tools to help understand microbiome-host interaction

Toot Suite — Prediction Overview



Predictors

Toot-SC — substrate
Toot-TC — TC info
Toot-All — all classifications

Toot-Proteome predict classification for membrane protein in a proteome, or meta-proteome

Toot-SS specific substrate for transport protein

Experimental Platform

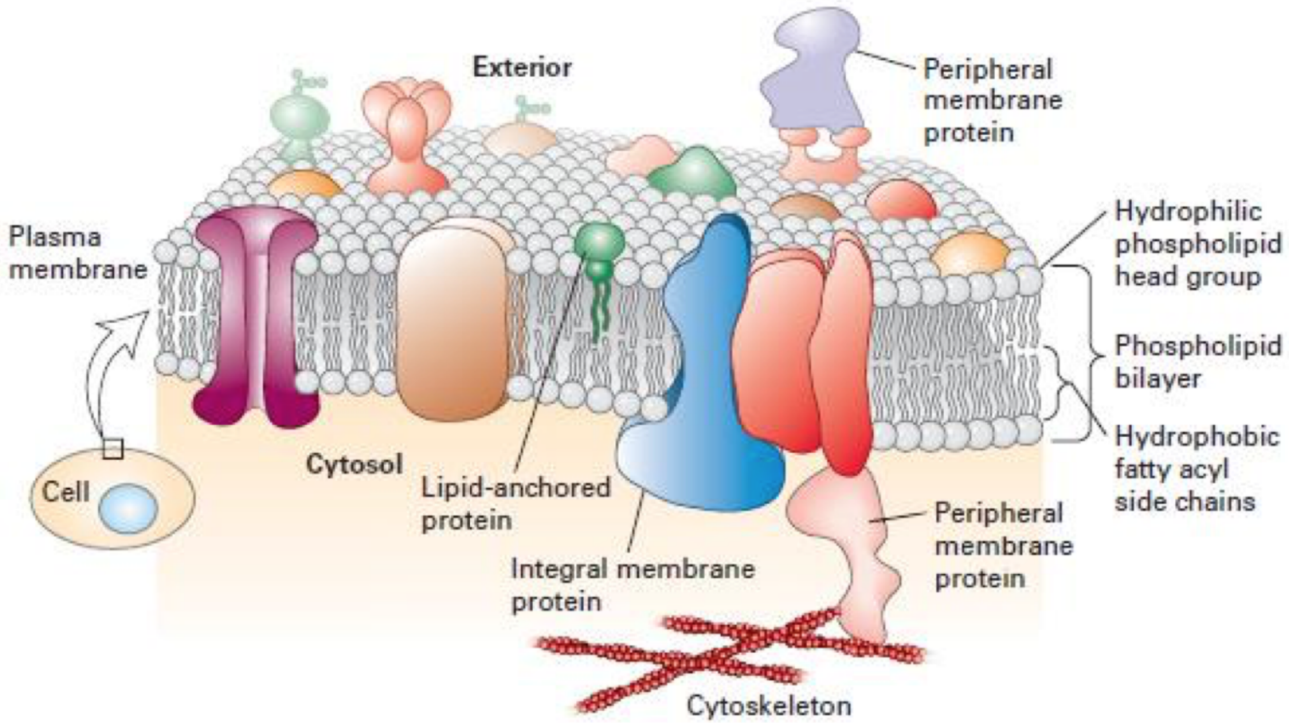
Experiments

TooT Suite Project 24 Substrate Classes

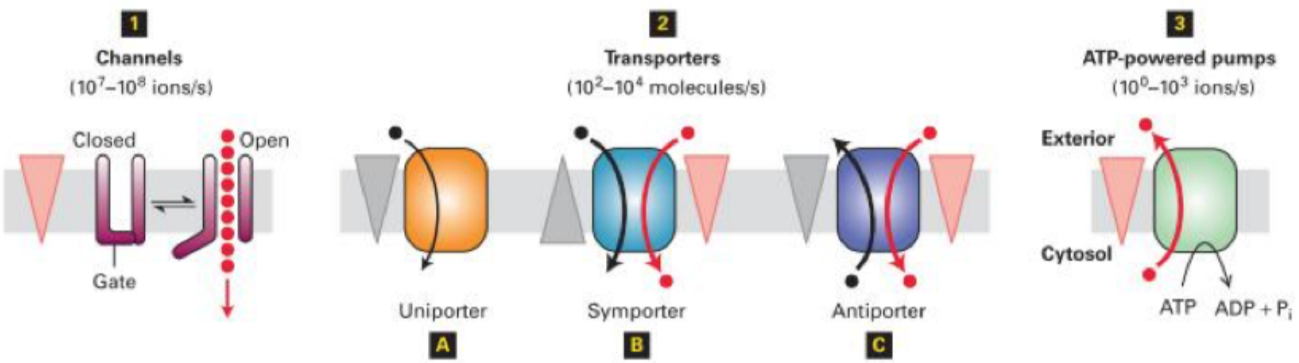
Table 5: Substrate Classes
Non-selective ions
Cations
Anions
Electrons
Water
Sugar and polyols
Monocarboxylates
Di- and tri-carboxylates
Organo-anions
Aromatic compounds
Amino acids and conjugates
Amines, amides, polyamines, and organo-cations
Siderophores, siderophore-Fe complexes
Substrate co-factors
Multiple drugs
Specific drugs
Other hydrophobic substrates
Nucleobases
Nucleosides
Polysaccharides
Proteins
Lipids
Nucleic acids
Unknown

Transport Proteins

Biomembrane



Transmembrane Transport Proteins



Membrane Proteins

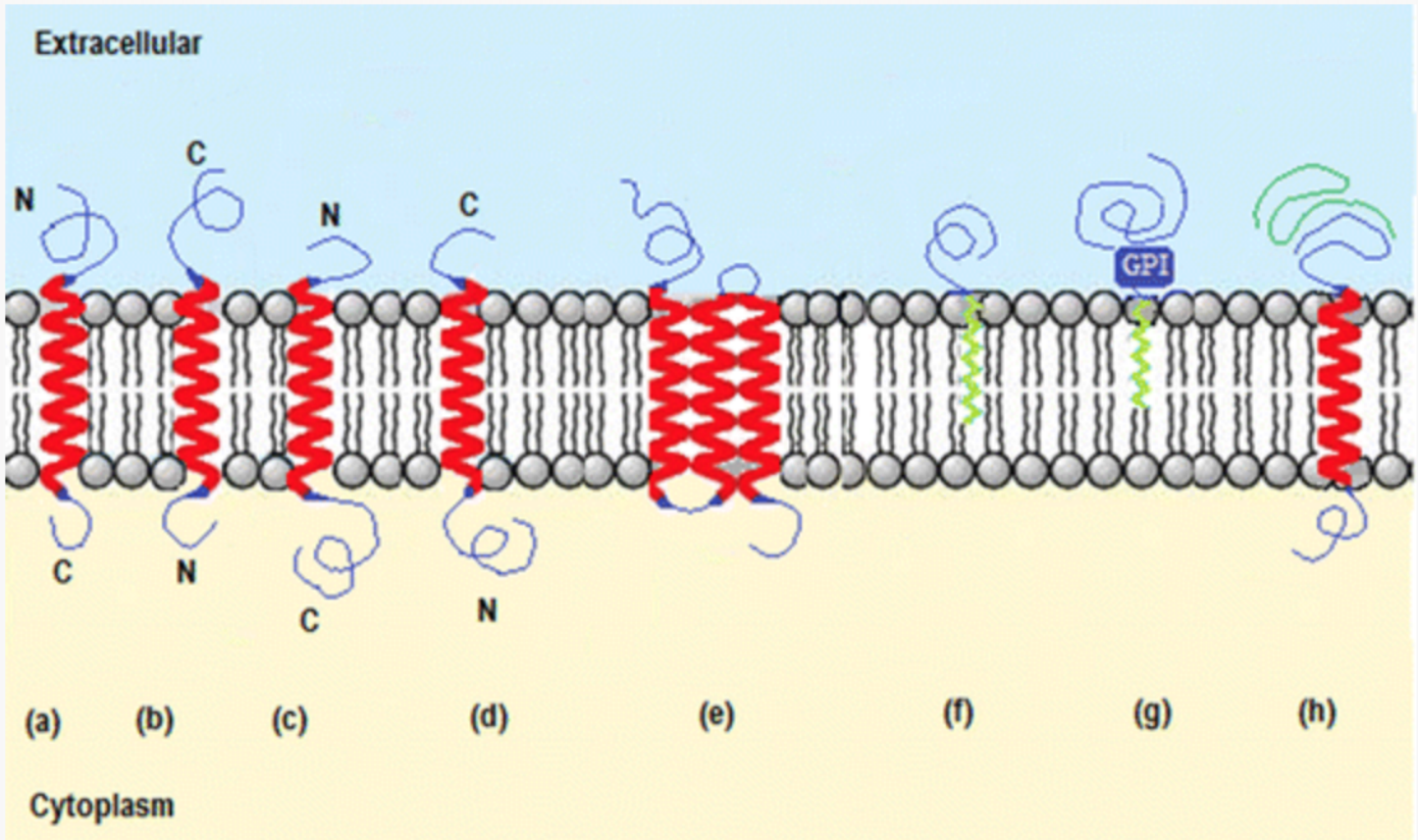
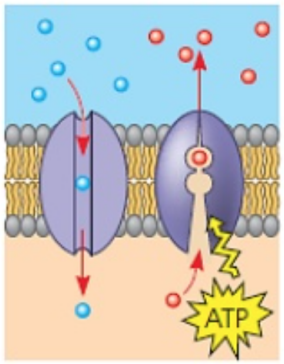


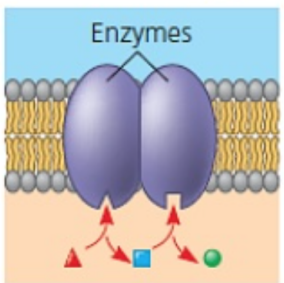
Fig. 2
Eight types of membrane protein with extracellular and intracellular activities. *a* Single-pass Type I. *b* Type II. *c* Type III. *d* Type IV. *e* Multi-pass transmembrane. *f* Lipid-anchored. *g* GPI-anchored. *h* Peripheral membrane proteins

Membrane Proteins

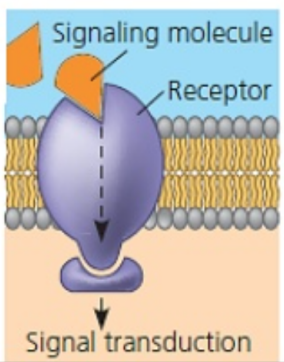
(a) Transport. *Left:* A protein that spans the membrane may provide a hydrophilic channel across the membrane that is selective for a particular solute. *Right:* Other transport proteins shuttle a substance from one side to the other by changing shape (see Figure 7.17). Some of these proteins hydrolyze ATP as an energy source to actively pump substances across the membrane.



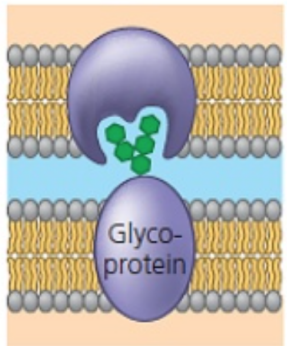
(b) Enzymatic activity. A protein built into the membrane may be an enzyme with its active site exposed to substances in the adjacent solution. In some cases, several enzymes in a membrane are organized as a team that carries out sequential steps of a metabolic pathway.



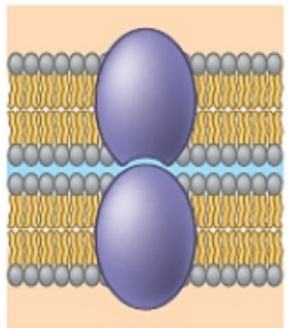
(c) Signal transduction. A membrane protein (receptor) may have a binding site with a specific shape that fits the shape of a chemical messenger, such as a hormone. The external messenger (signaling molecule) may cause the protein to change shape, allowing it to relay the message to the inside of the cell, usually by binding to a cytoplasmic protein (see Figure 11.6).



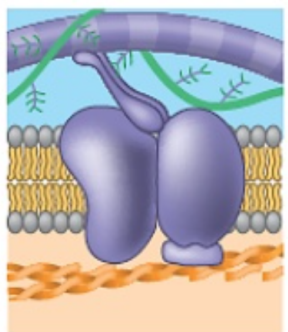
(d) Cell-cell recognition. Some glycoproteins serve as identification tags that are specifically recognized by membrane proteins of other cells. This type of cell-cell binding is usually short-lived compared to that shown in (e).



(e) Intercellular joining. Membrane proteins of adjacent cells may hook together in various kinds of junctions, such as gap junctions or tight junctions (see Figure 6.32). This type of binding is more long-lasting than that shown in (d).



(f) Attachment to the cytoskeleton and extracellular matrix (ECM). Microfilaments or other elements of the cytoskeleton may be noncovalently bound to membrane proteins, a function that helps maintain cell shape and stabilizes the location of certain membrane proteins. Proteins that can bind to ECM molecules can coordinate extracellular and intracellular changes (see Figure 6.30).



19

Campbell *et al*, Biology; 2009.

Toot Suite — Scale

Membrane proteins

Proteome

Meta-proteome

Machine Learning Experiments

EPRCS Methodology

- Evolutionary information
- Positional information
- Regional information
- Compositional information
- Sequential information

Experiments

What is the

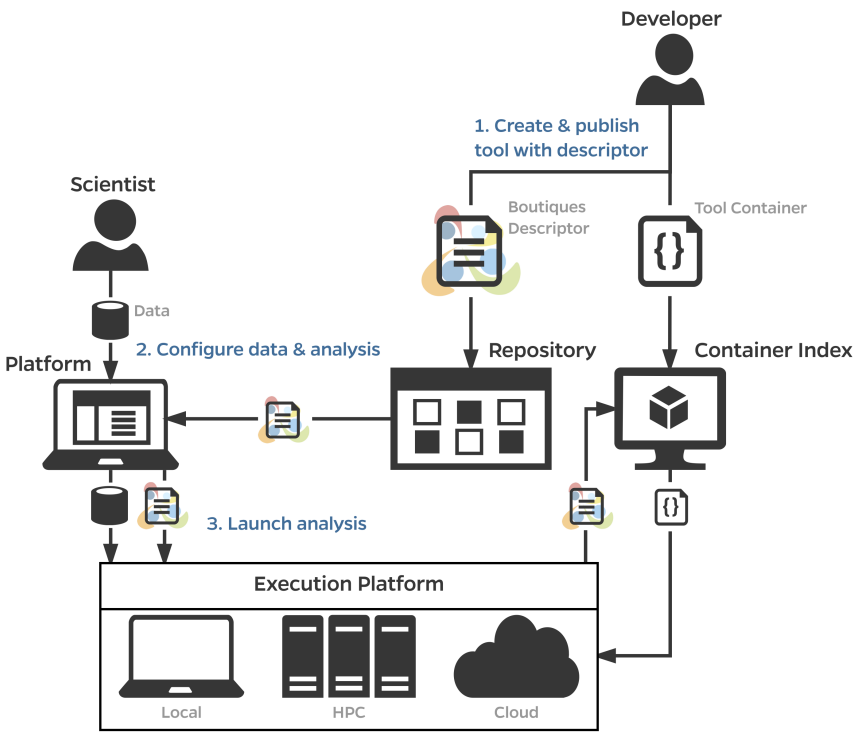
best combination of

E P R C S tools

for prediction

Toot Suite — Experimental Infrastructure

Boutiques using Docker



Compute Canada
e.g. MP2 cluster

1632 nodes
12 core/node
32-512 GB memory/node

FAIR for Open Science

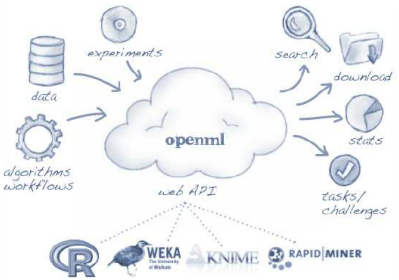
Findable
Accessible
Interoperable
Reusable

T Glatard et al, Boutiques: a flexible framework to integrate command-line applications in computing platforms.

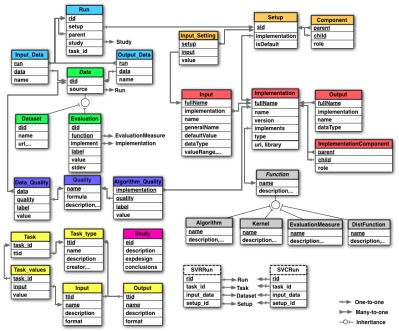
Gigascience. 2018 May 1;7(5)

Toot Suite — F.A.I.R.

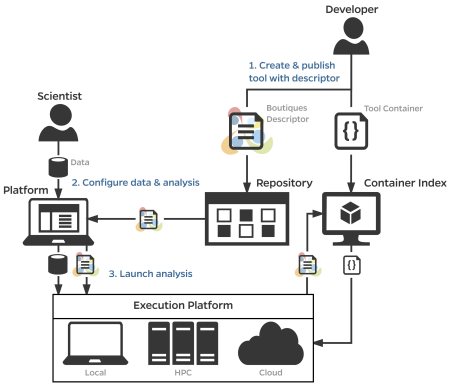
openML



openML



Boutiques



Data Sources

- UniProt
- TCDB
- GO, ChEBI, Pfam
- ABC, SLC

Findable

Zenodo, doi

Accessible

DockerHub

Interoperable

Boutiques, Galaxy, CWL

Reusable

Boutiques, openML

MD Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016

Outline

EPRCS Methodology

- EPRCS — Evolution
- EPRCS — Position
- EPRCS — Region
- EPRCS — Composition
- EPRCS — Sequence
- EPRCS — Example TranCEP

EPRCS Methodology for Protein Sequence Analysis

Evolution **[E]**

Classical `blastp`, `PSI-blast`
MSA, TMS-aware MSA

Position **[P]**

Focus on important sites
classical PSSM

Region **[R]**

Split sequence into regions
eg C-terminus, Rest, N-terminus
eg TMS and non-TMS

Composition **[C]**

Classical amino acid composition
AAC, PAAC, PseAAC (Chou), split

Sequence **[S]**

HMM capture patterns along sequence

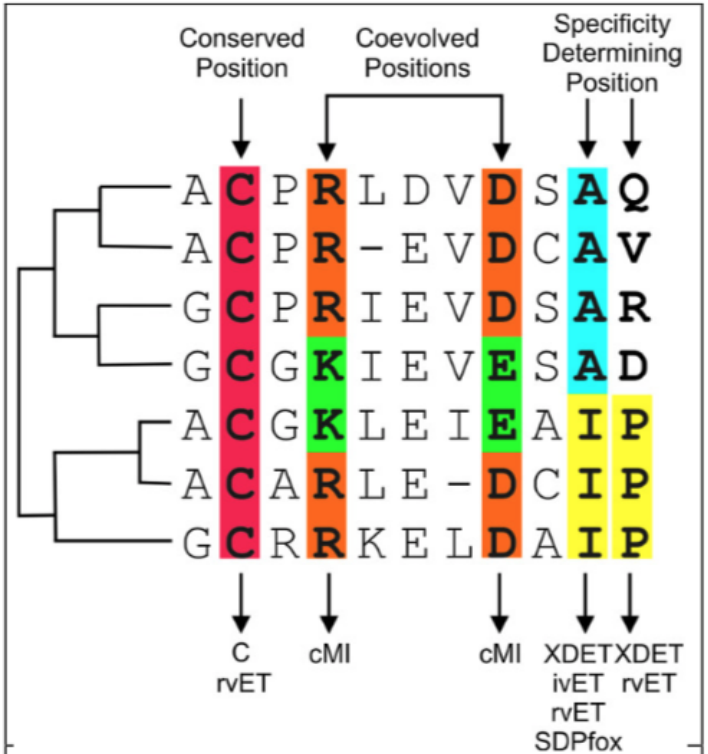


Figure 4: Types of Positional Information [Teppa'12]

TranCEP — Predicting transport proteins

Algorithm

Construct vector *v(s)* for protein sequence *s*

- **[E]** Form set *S* of similar sequences to *s* in DB using `blastp`
- **[E]** Form MSA *M* from *S* using `TM-Coffee`
- **[P]** Find informative positions *P* in *M* using `TCS`
- **[P]** Filter uninformative positions in *M* to form *m*
- **[C]** Compute PAAC composition vector *v(s)* from *m*

TranCEP builds 21 SVMs

discriminating 1-vs-1 for each pair of the 7 classes

Munira Alballa, Faizah Aplop, Gregory Butler, *TranCEP: Predicting transmembrane transport proteins using composition, evolutionary, and positional information*, bioXriv, 2018. doi: https://doi.org/10.1101/293159

TranCEP — Predicting transport proteins

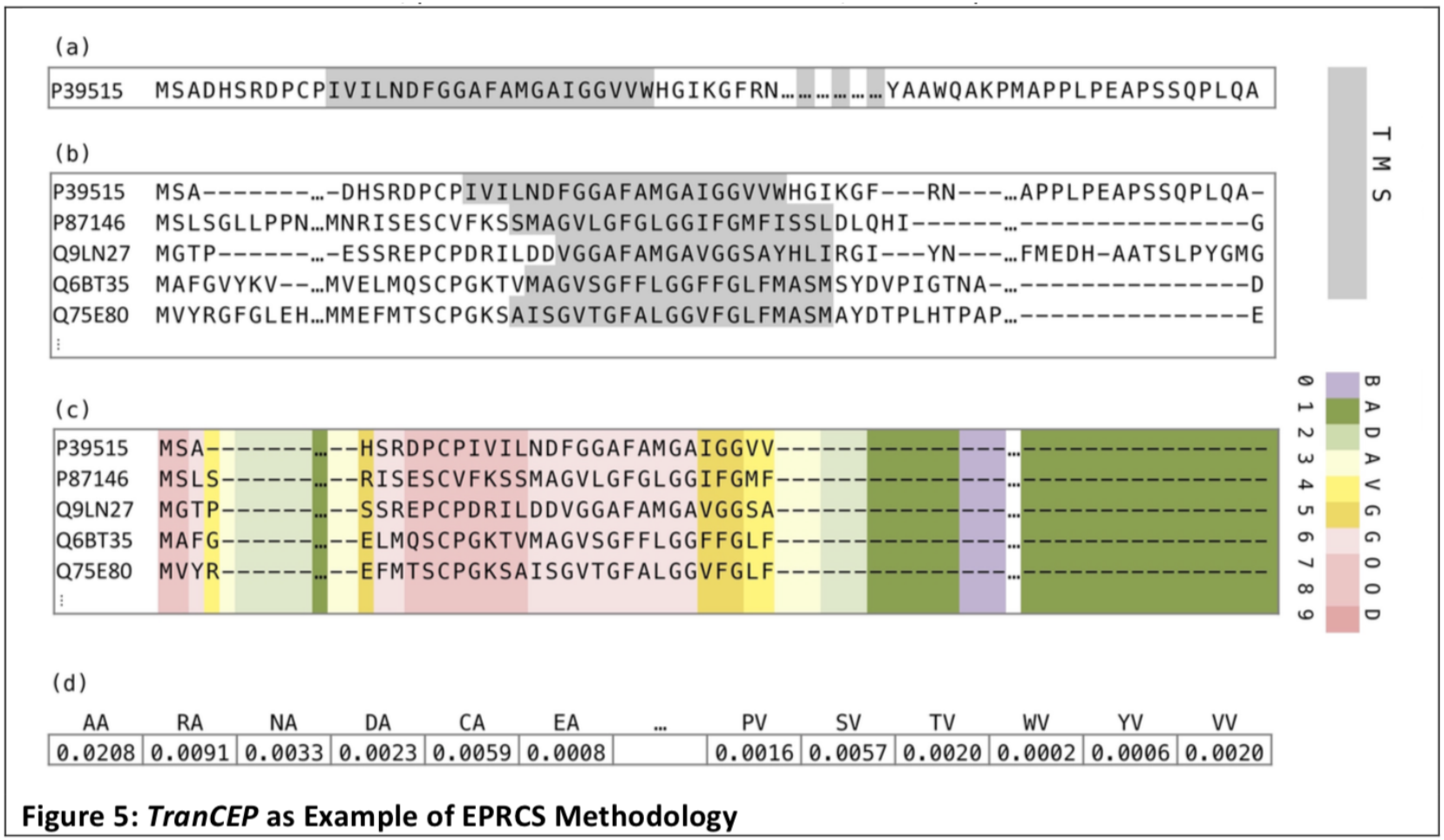


Figure 5: *TranCEP* as Example of EPRCS Methodology

TranCEP — Predicting transport proteins

Performance of TranCEP using TM-Coffee and TCS and PAAC

Class	Specificity		Sensitivity		Accuracy		MCC	
	TrSSP	TranCEP	TrSSP	TranCEP	TrSSP	TranCEP	TrSSP	TranCEP
Amino acid	82.42	98.10	93.33	60.00	83.33	91.75	0.49	**0.66**
Anion	69.05	96.30	75.00	58.33	69.44	90.82	0.23	**0.56**
Cation	74.31	89.29	75.00	94.44	74.44	89.00	0.41	**0.78**
Electron	91.78	99.05	80.00	80.00	91.11	97.80	0.50	**0.88**
Protein	82.42	99.07	93.33	66.67	83.33	93.68	0.49	**0.75**
Sugar	76.79	99.07	91.67	66.67	77.78	94.68	0.38	**0.74**
Other	73.13	86.00	60.00	65.00	71.67	80.91	0.23	**0.44**
Overall					78.88	74.17	0.41	**0.69**

Table 3: Comparison of *TranCEP* and *TrSSP*

TportHMM

TportHMM is MUSCLE plus Xdet plus `hmmer`
MCC of 0.72 on Mishra's dataset

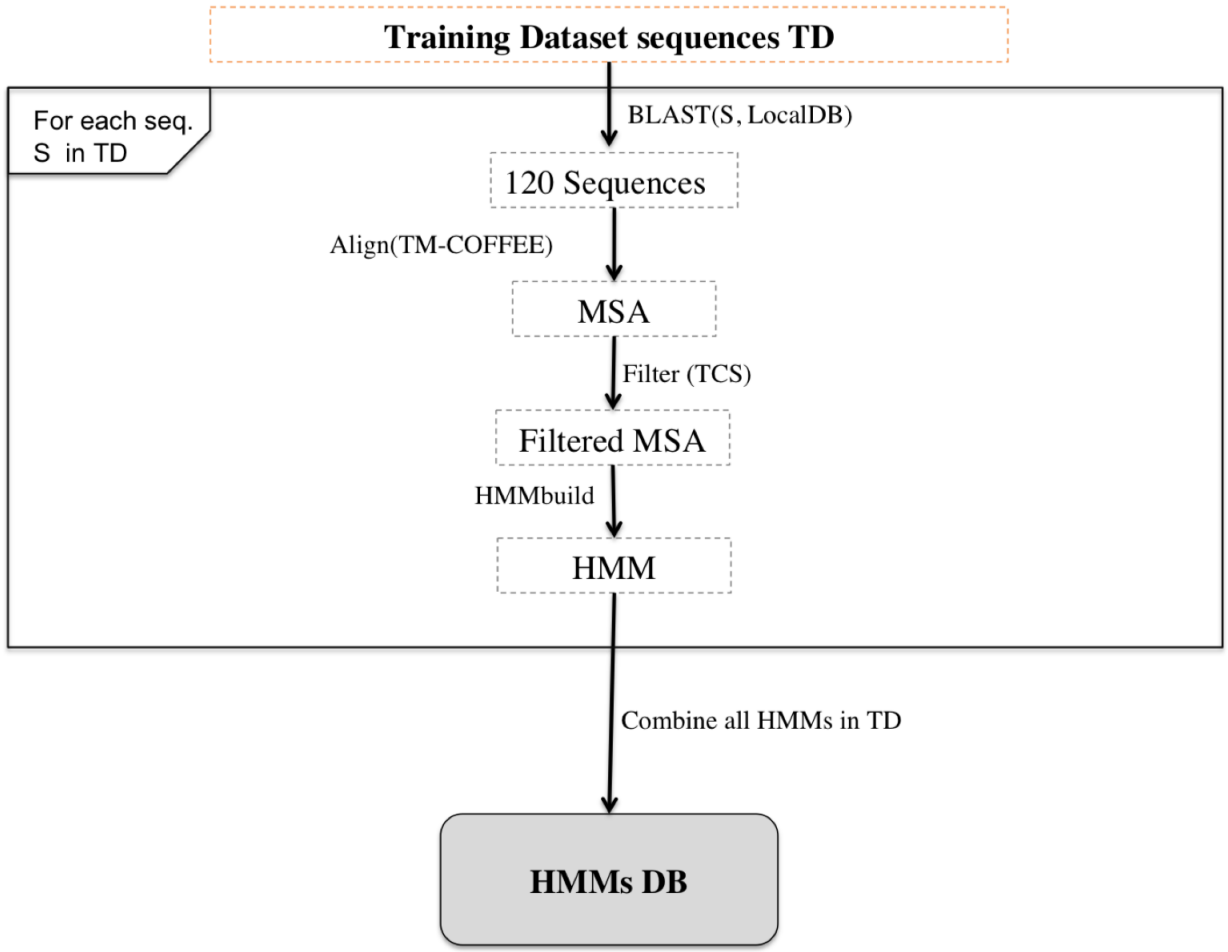


Figure 5: Filtered-HMM-Profile database building process

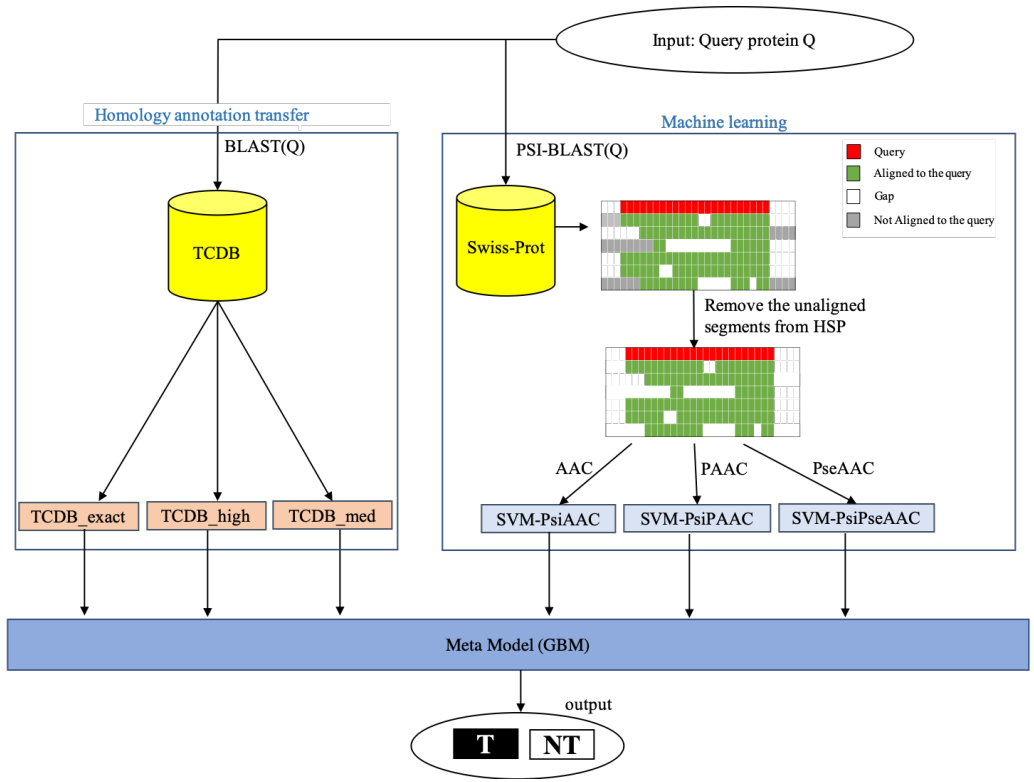
TooT-SC: TM-Coffee+PAAC+SVM on 11 classes

MCC 0.79

	Classs name	ChEBI-ID	#instances	Main decentents
1	nonselective inorganic molecule	(CHEBI:36914 - inorganic ion) and (CHEBI:24431 - chemical entity)	26	
2	water	CHEBI:15377 - water	26	
3	inorganic cation	CHEBI:36915 - inorganic cation	601	
4	inorganic anion	CHEBI:24834 - inorganic anion	102	
5	organic anion	CHEBI:25696 - organic anion	107	
6	organooxogyn	CHEBI:36963 - organooxygen compound	174	
				CHEBI:35381;CHEBI:63367 - monosaccha
				CHEBI:50699; CHEBI:63563 - oligosaccha
				CHEBI:18154; CHEBI:65212 -polysacchari
				CHEBI:25384 -monocarboxylic acid -
				CHEBI:35692 -dicarboxylic acid
				CHEBI:27093 -tricarboxylic acid
7	amino acid and derivatives	(CHEBI:33709 - amino acid) and (CHEBI:83821 - amino acid derivative)	157	
8	other organonitrogen compound		160	
				CHEBI:16670 - peptide
				CHEBI:32952 - amine
				CHEBI:88061 - polyamine
				CHEBI:36080 - protein
				CHEBI:50047 - organic amino compound
9	nucleotide	CHEBI:36976 - nucleotide	24	
10	organic heterocyclic		37	
				CHEBI:18282 - nucleobase
				CHEBI:33838 - nucleoside
				CHEBI:33696 - nucleic acid
11	miscellaneous		110	
				CHEBI:26191 - polyol
				CHEBI:25703 - organic phosphate
				CHEBI:32988 - amide
				CHEBI:50860 - organic molecular entity
				CHEBI:25697 - organic cation

TooT-T : Discrimination of transport proteins from non-transport proteins

Munira Alballa[1*] and Gregory Butler[1,2]



Conclusion: The proposed model outperforms all of the state-of-the-art methods that rely on the protein sequence alone, with respect to accuracy and MCC. TooT-T achieved an overall accuracy of 90.07 % and 92.22% and an MCC 0.80 and 0.82 with the training and independent datasets, respectively.

Outline

Conclusion — Challenges

Sharing our ML experiments

openML excellent for sharing ML experiments
automated parameter tuning and hyperparameter tuning

openML handles tabular data, not sequence data
issues with categorical data (from python scikit-learn)
big issue with heirarchical categorical data

Performance of Classification Step

MSA and SDS may be time-consuming
Need to classify 10M proteins for microbiomes

Predicting transport of specific substrates

are there enough examples known
multi-class vs multi-label learning

Thank You!

Questions, Please?