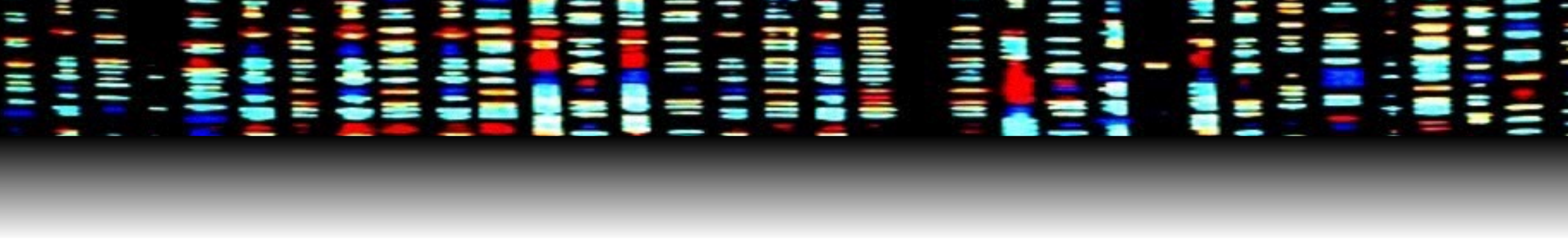


Analytical Systems for Life Sciences at the time of Big Data

Fabrizio Costa

Freiburg University

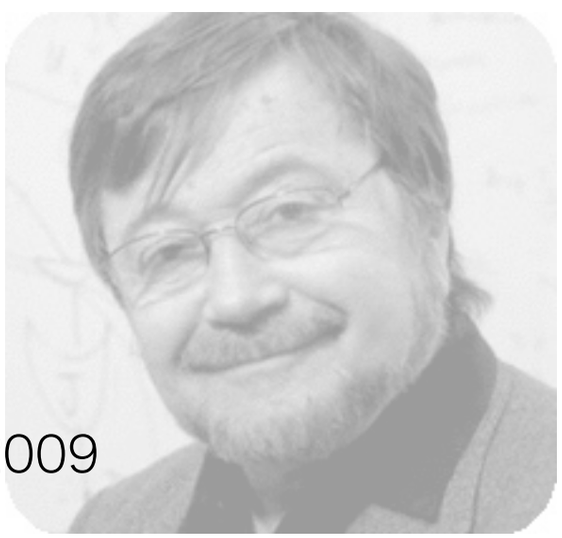
Motivation

How can analytical approaches exploit increasingly larger amount of heterogeneous data to help scientific knowledge acquisition?

The human in the loop

- Ultimately researchers want analytical models in order improve:
  - understanding
  - actionability

- **Note:**
  interactive processing is key
  for human intuitions



Pearl, J. Causal inference in statistics: An overview. Stat. Surv. 2009

Data quantity

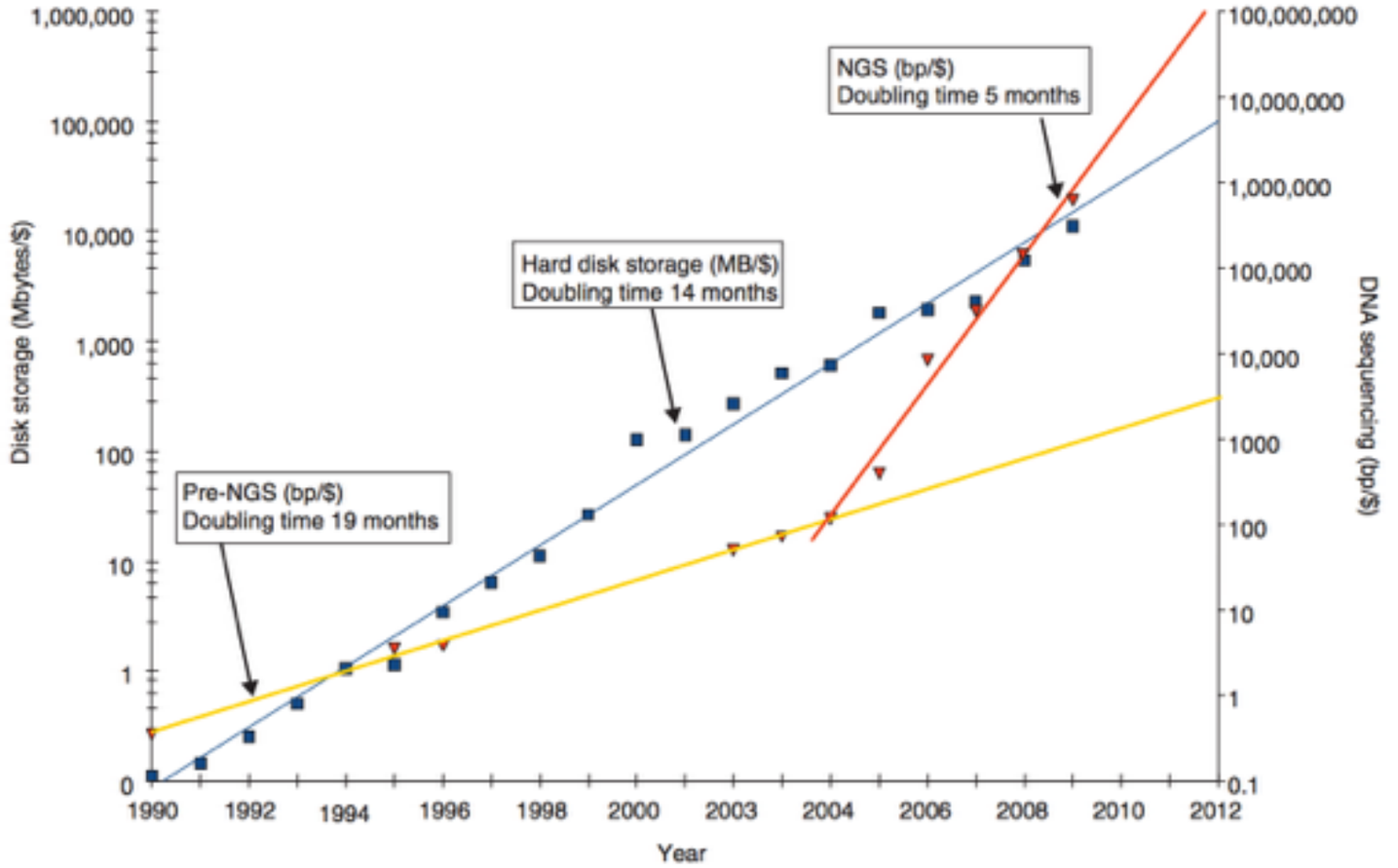


Image from Stein, Lincoln D. 2010. Genome Biology 11 (5): 207.

Performance doubling time (in months):
cpu:18　disk:13　transfer:09　sequencing:05

cpu: Moore, Gordon E. 1965 Electronics
disk: Walter, Chip. 2005 Scientific American
transfer: Reynolds, Carson. 1998 ACM SIGCHI Bulletin
seq: Stein, Lincoln D. 2010 Genome Biology

Data complexity

- We can now know in high-throughput (and in vivo):

1. the identity of entire classes of biological entities
   - RNAs [RNA-Seq], proteins, peptides and metabolites [mass spectrometry]
2. … and their relationships (interactions)
   - prot-prot [yeast two-hybrid], prot-RNA [iCLIP], RNA-RNA [CLASH]

yeast two-hybrid: Luban, J. & Goff, S. P. Curr. Opin. Biotechnol. 1995
RNA-Seq: Wang, Z., et al. Nat. Rev. Genet. 2009
iCLIP: Huppertz, I. et al. Methods 2014
CLASH: Kudla, G., et al. PNAS 2011

Desiderata

- Given the data explosion in quantity and complexity we need analytical systems that exhibit:

1. adaptive bias

2. efficiency <-> simplicity
   - computational viewpoint: efficiency
   - human viewpoint: simplicity (abstractions)

Approaches

- Instead of committing to a specific ML approach
  connectionist approach vs symbolic approach vs Bayesian approach vs …

- **generic** computational approach
  structured composition of parametrized objects that can do certain classes of computations (in some consistent way) and whose parametric configuration can be adapted (in some useful way) to a context or a task

- The driving design force should be efficiency coupled with simplifying abstraction management not a specific computational paradigm

Algebraic paradigm

- Towards and ‘algebraic’ paradigm in ML

  Bottou, L. From machine learning to machine reasoning. Mach. Learn. 2014

- Algebra: collection of operations closed on a specific domain

  Reasoning as computation, possibly of many kinds

- **Question:** in order to develop general purpose analytical systems for life sciences

  1. which (flexible) domain (data type)?
  2. which (few) operations?

Domain

- iterators ADT over graphs ADT
- Advantages of graph formalism:

1. universal encoding
   - state of computation can be encoded in the graph itself
2. intuitive for humans
   - conceptual entities are nodes and relations are edges
   - nesting graphs can represent abstractions

Processing

- Ideas from Generic Programming and Abstract Data Types:
  - abstract solutions to specific class of problems
    - encapsulation/abstraction
    - localization/interface
    - flexibility/equivalence
- Ideas from Functional Programming:
  - referential transparency
  - compositionality

Generic Computational Algebra

operation(iterable, program, priors, precond, postcond)

- operation declares the abstract type of action/problem
- program declares the algorithm for the solution
- parameters_priors declares the user prior knowledge on the program's parameters' space
- precondition/postcondition declares the conditions to be fulfilled on input/output for the pair (operation,program)

Generic operations

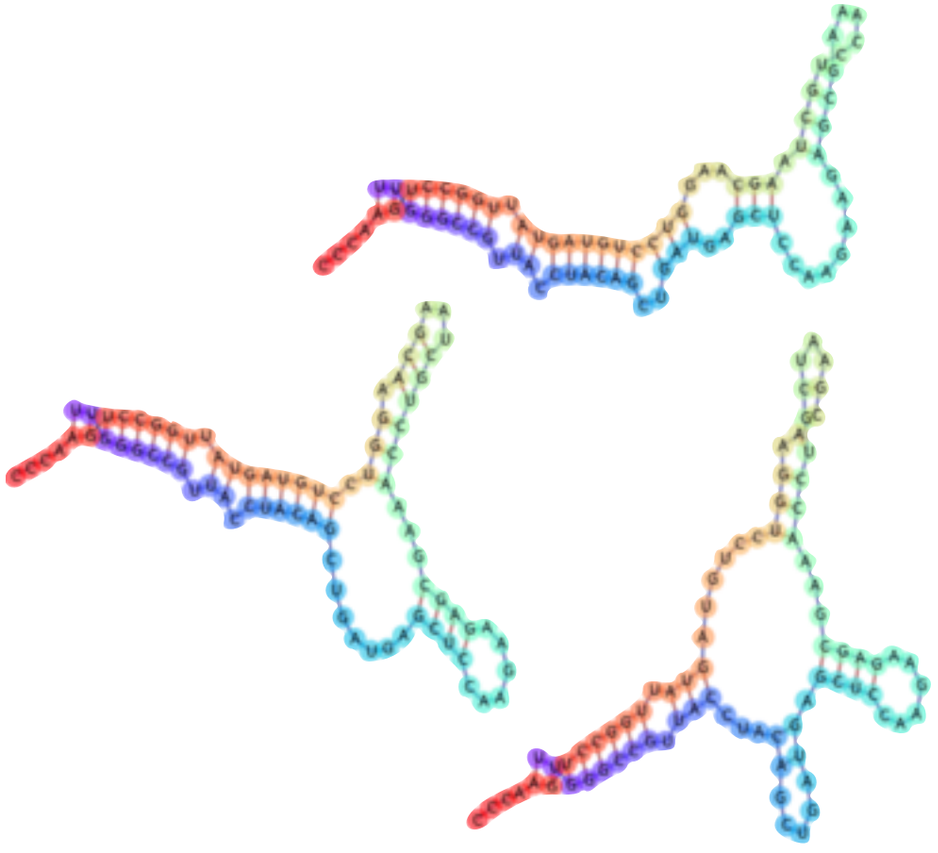
iterator over * > iterator over *

1. convert: any type > graph
2. associate: graph > any type
3. partition: graphs > iterators over graphs
4. decompose: graph > iterator over (sub)graphs
5. compose: iterator over graphs > graph’
6. transform: graph > graph’
7. order: graph > graph
8. construct: graph > graphs’

Convert

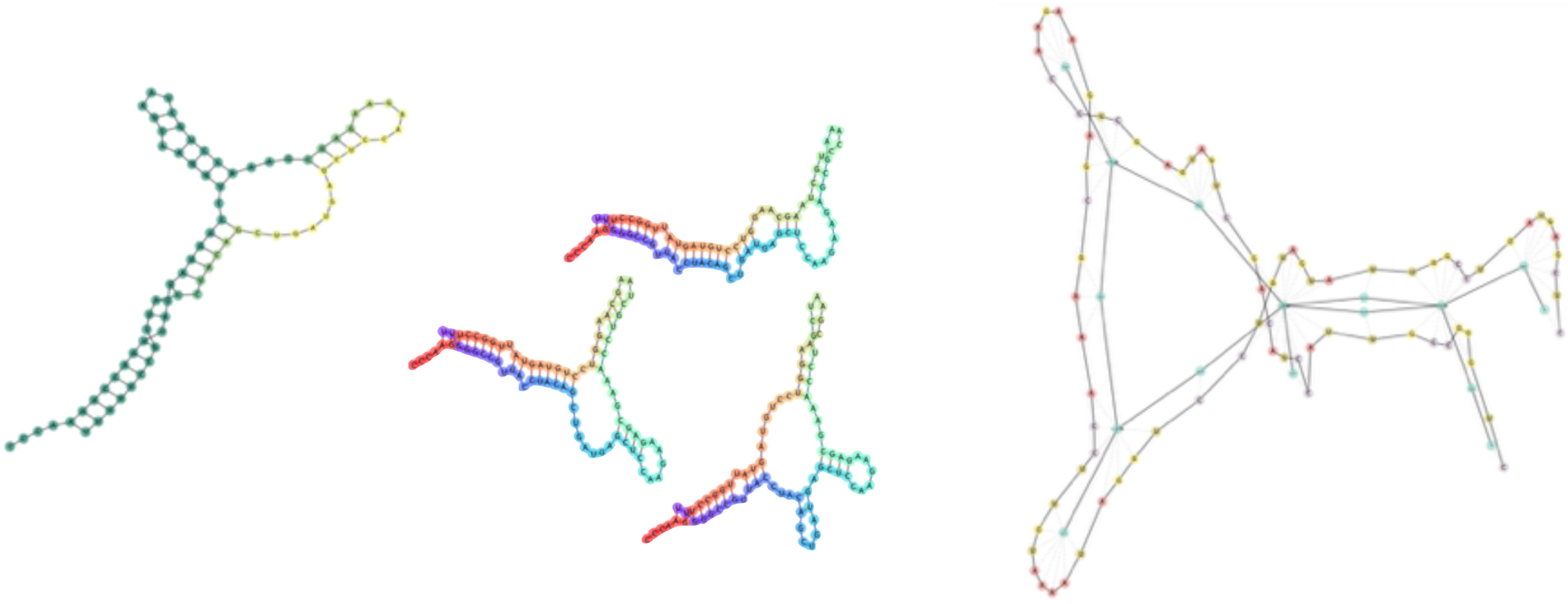
- Ex: build graph encoding from other data types

```
>ABQF01059171.1/305-384
UUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUGCUAAGCAAGGUCC
UGUAGUAUUGGCCUGAACCC
>AADN03003451.1/4511-4593
CUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUGUAAAAUAGGUCC
UGUAGUAUUGGCCUGAUGAGCUC
>AAWZ02032198.1/15823-15741
UGAGGCCGUUACCUACAGCUGAUGAGCUCCAAAAAGAGCGAAACCUGUUAAAAUAGGUCC
UGUAGUAUUGGCCGACUGAGCCG
>AGAI01055016.1/63287-63205
UUAGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUUUAAGAUAGGUCC
UGUAGUAUUGGCCUGAAAACCAU
>AANN01066007.1/588-511
CUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUGCUAGGUCCUGCAG
UACUGGCUUAAGAGGCUA
>AAQR03161315.1/4048-3972
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUAAGAGAAU
>ABRN01375670.1/21703-21777
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUGAGAUA
```



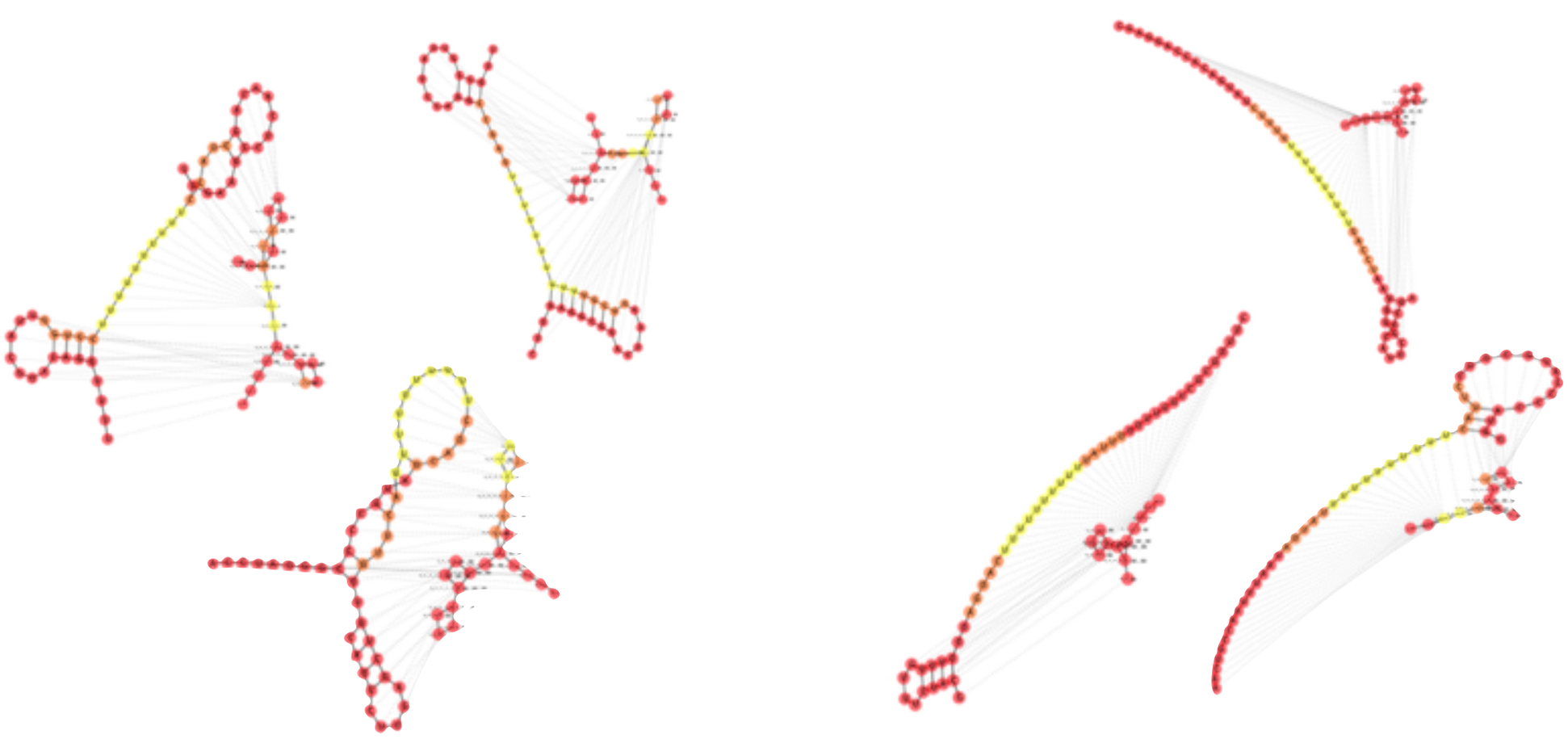
Associate

- Ex: supervised paradigm
  discover which hypothesis are likely to improve associability (i.e. predictability)



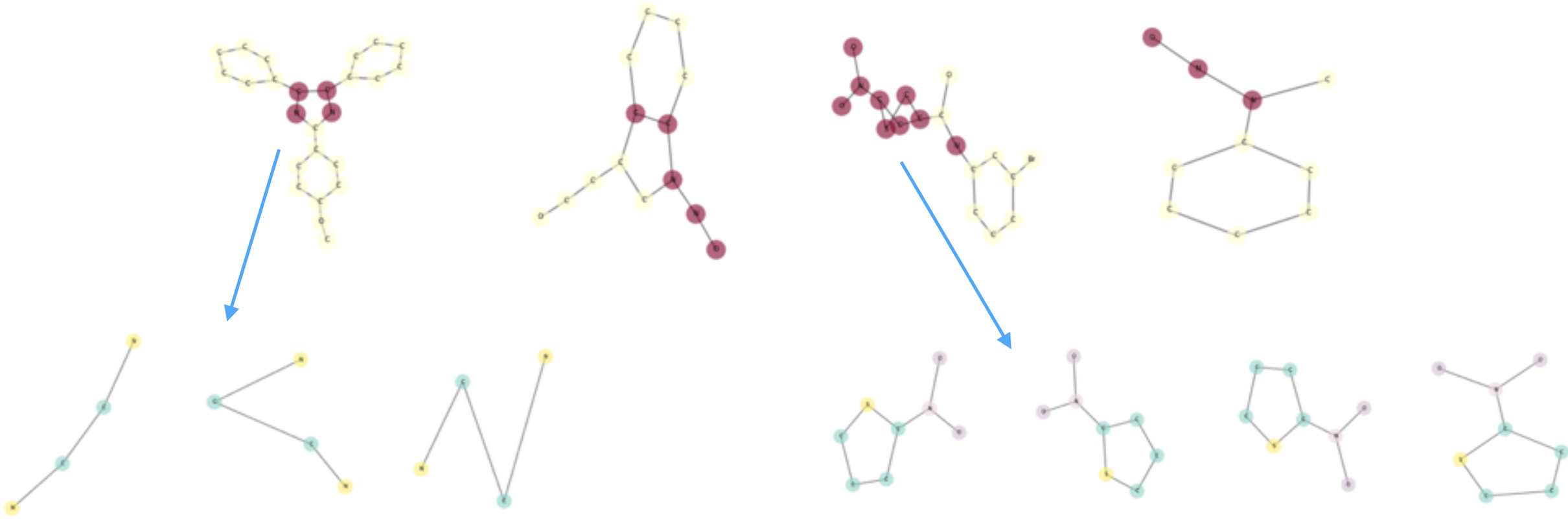
Partition

- Ex: find structure in collections of instances



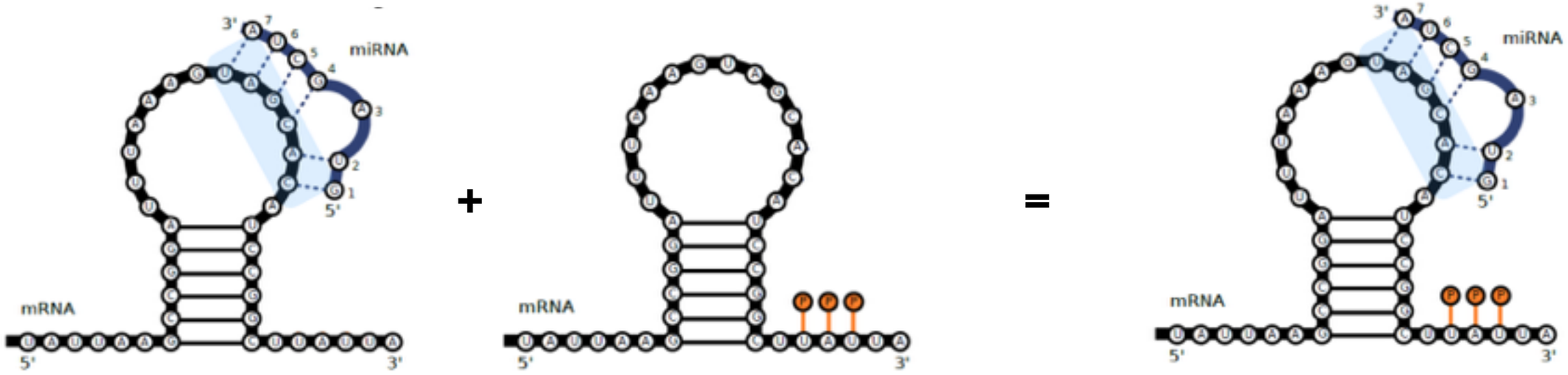
Decompose

- Ex: find structure in parts of instances



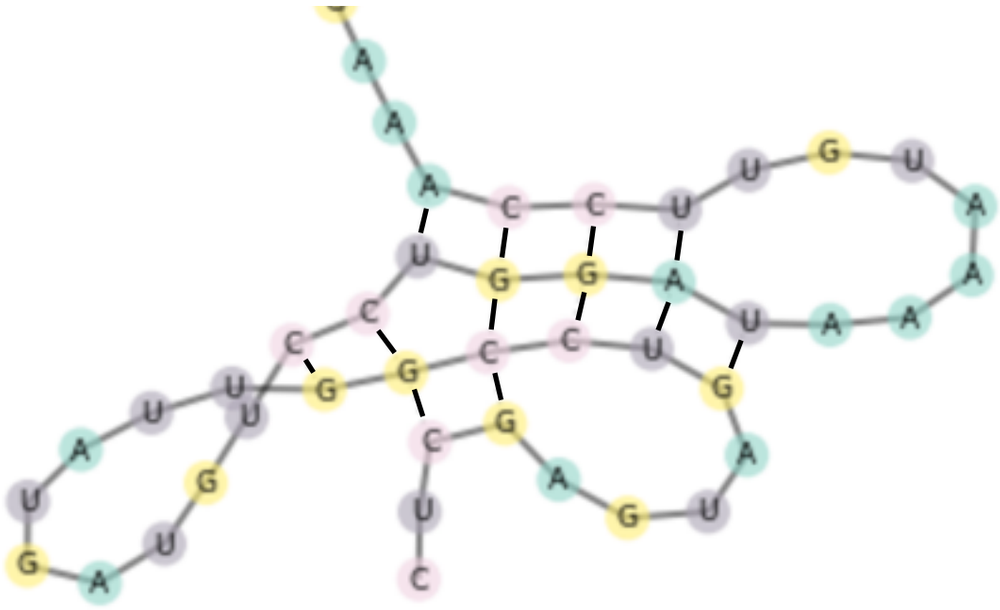
Compose

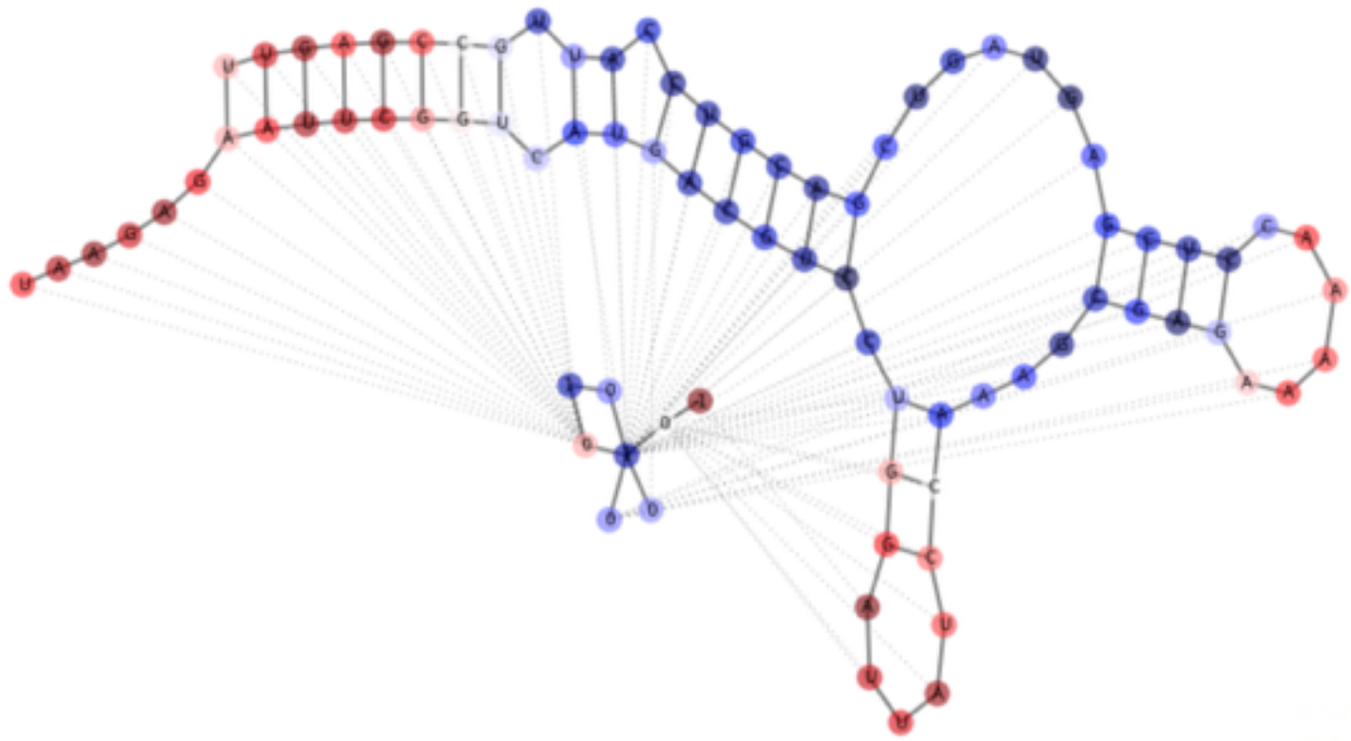
- Ex: combine two or more instances into a graph



Transform

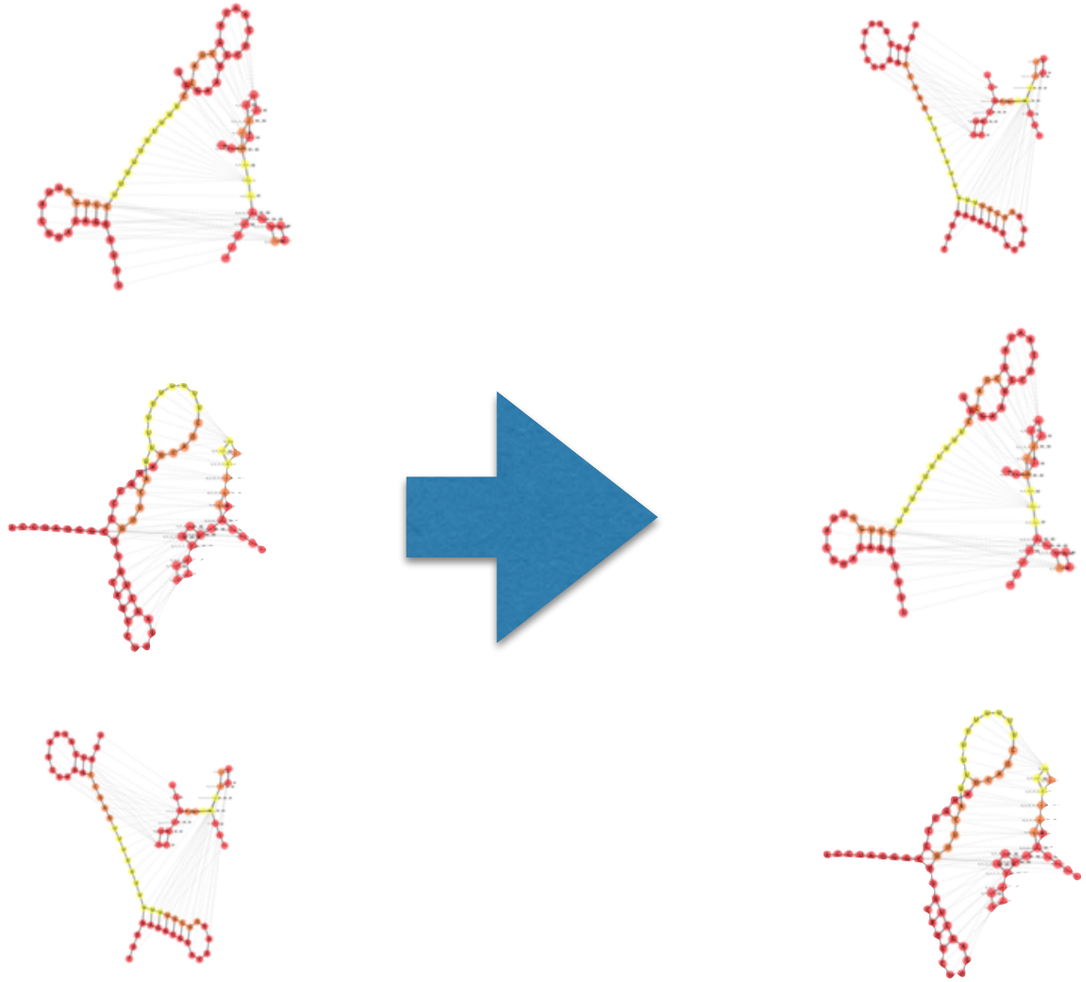
- Ex: change representation (!)
  possibly using previously adapted systems





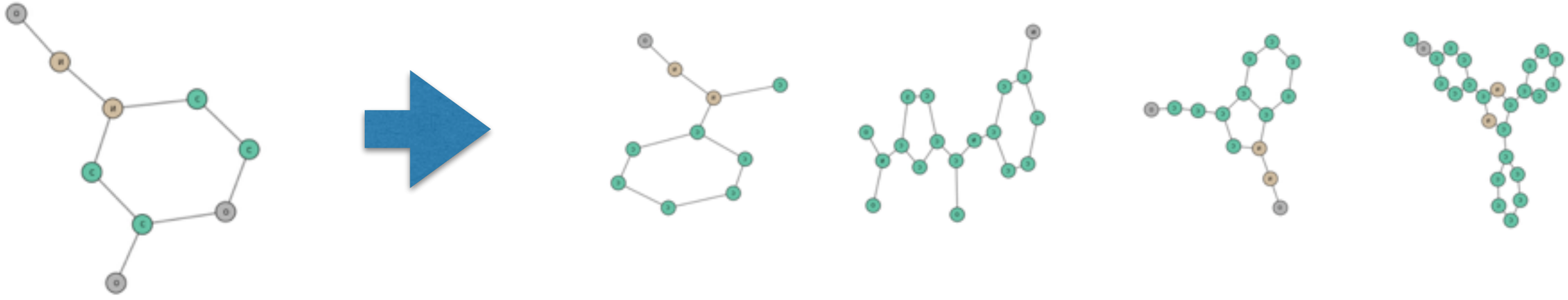
Order

- Ex: re-arrange the order of graphs based on their representativeness



Construct

- Ex: build data with desired properties



```
>ABQF01059171.1/305-384
UUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUGCUAAGCAAGGUCC
UGUAGUAUUGGCCUGAACCC
>AADN03003451.1/4511-4593
CUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUGUAAAAUAGGUCC
UGUAGUAUUGGCCUGAUGAGCUC
```

```
>A1
UUAGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUUUAAGAUAGGUCC
UGUAGUAUUGGCCUGAAAACCAU
>A 2
CUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUGCUAGGUCCUGCAG
UACUGGCUUAAGAGGCUA
>A 3
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUAAGAGAAU
>A 4
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUGAGAUA
```

Implementation

1. Instances are encoded as weighted graphs with richly typed nodes and edges
   (extension to hyper graphs is possible)

2. Efficient mapping procedure graphs ▷ vectors
   traditional ML is directly applicable on resulting representation

3. Catalog of programs supporting proposed interface implemented on top of 1. & 2.

Mapping support

- Python library: EDeN

  Explicit Decomposition with Neighbourhoods

  evolution of NSPDK (Costa, De Grave ICML 2010)

1. fast mapping: near linear complexity

2. simple: exposes small/clear interface

   scikit-learn style

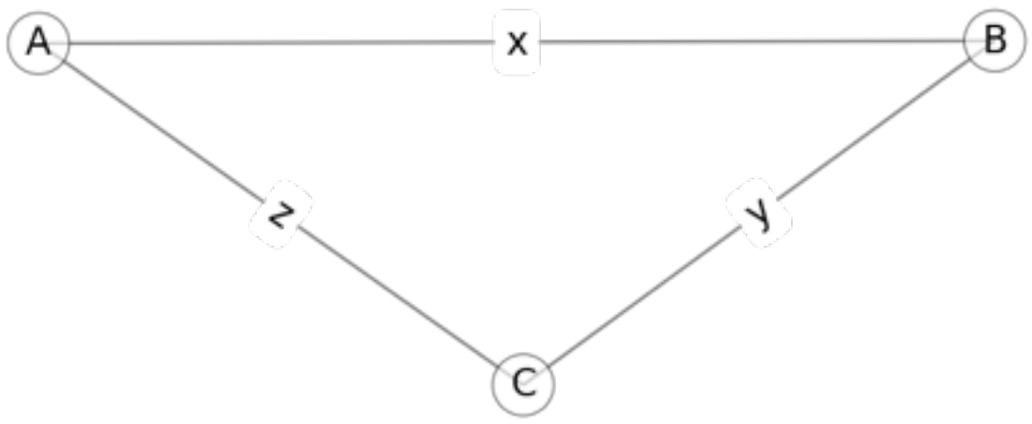
3. general purpose: heterogeneous graphs

EDeN

Explicit Decomposition with Neighbourhoods

pip install git+https://github.com/fabriziocosta/EDeN.git

- weighted graphs
- with labels on nodes and edges

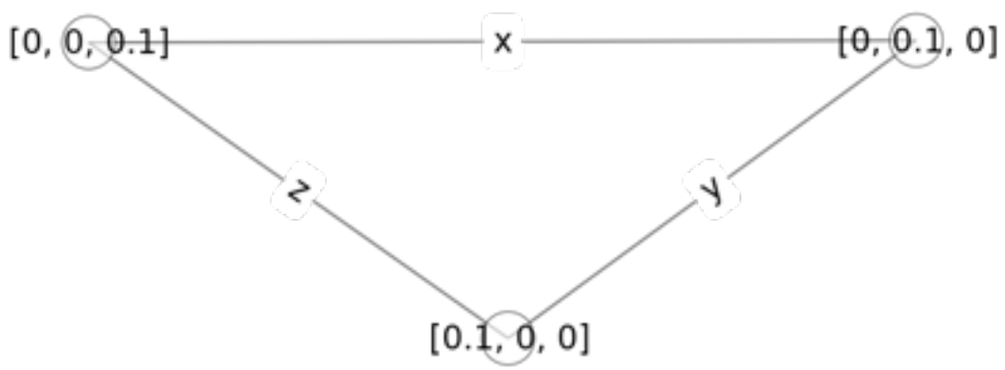


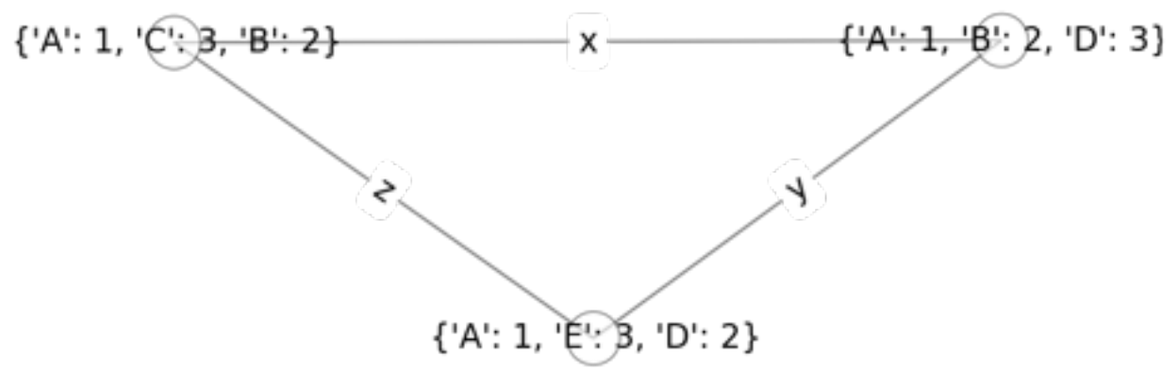
EDeN

Explicit Decomposition with Neighbourhoods

pip install git+https://github.com/fabriziocosta/EDeN.git

- labels can encode groups of reals as:
  - lists or dense vectors
  - dictionaries or sparse vectors





EDeN

Explicit Decomposition with Neighbourhoods

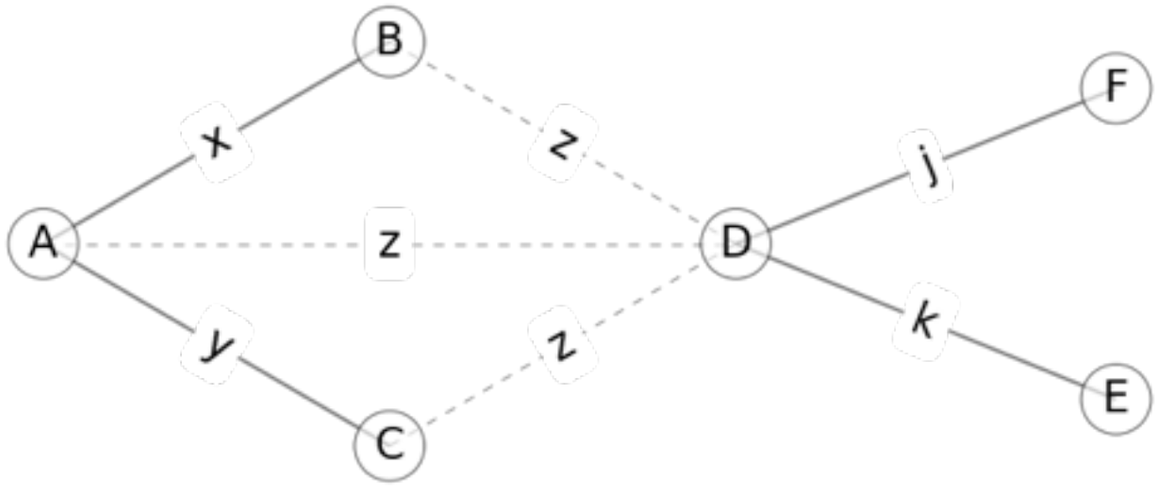
pip install git+https://github.com/fabriziocosta/EDeN.git

- graphs can be nested

  a nesting edge is a distinguishable type of edge

- nesting edges can represent abstractions like:

  part-of and is_a



Empirical run times graph to vector mapper

- instance mapping is perfectly parallelizable

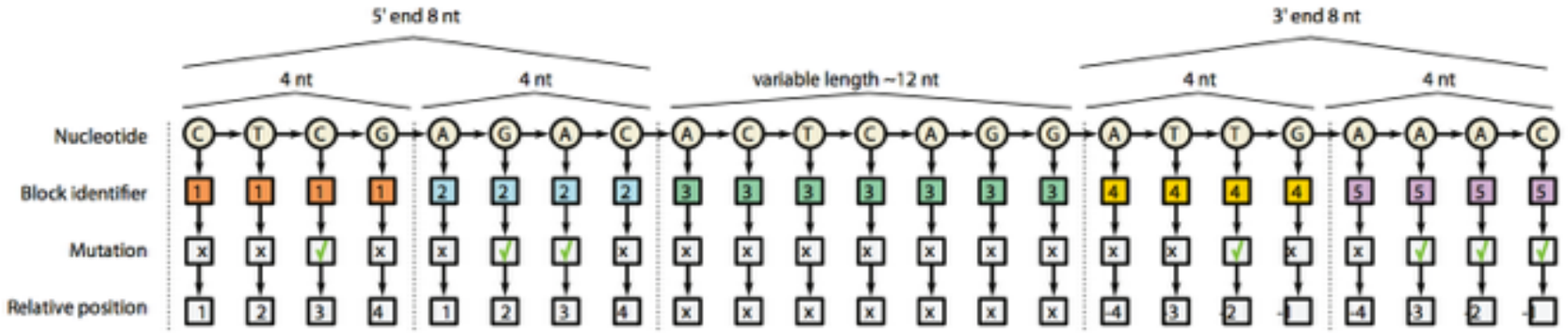
- molecules: 5000 graphs x min x core

- multi class prediction on RNA sequences

  3GHz machine 8 cores (C++ optimization for sequences)

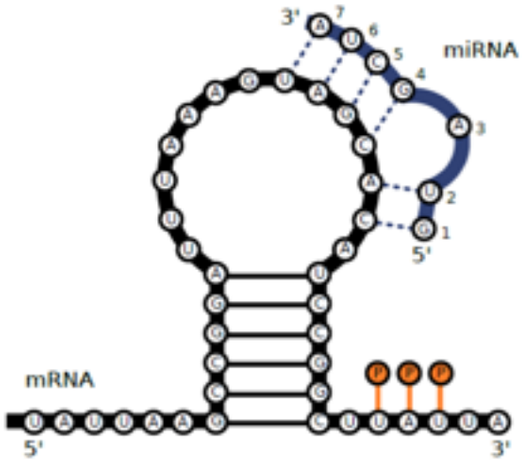
  - fit 1500 bacterial genomes: 8 min
  - predict metagenomic 32Mbases: 5 min

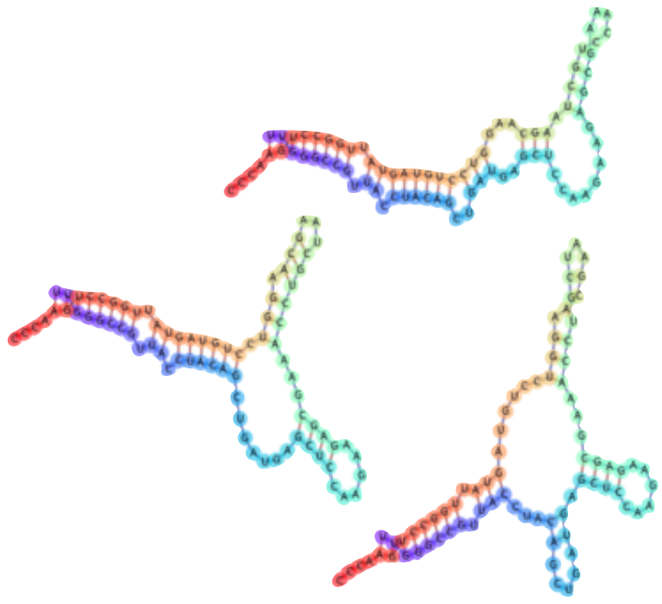
    (alignment based BLAST approaches take weeks on large cluster centers)

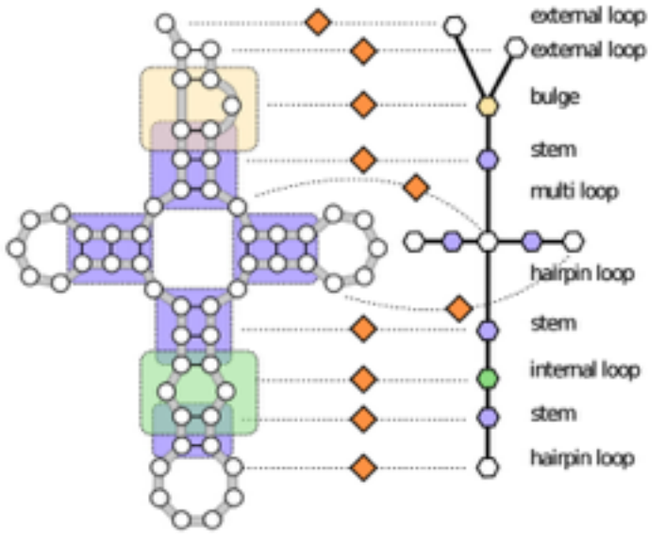


Application cases:

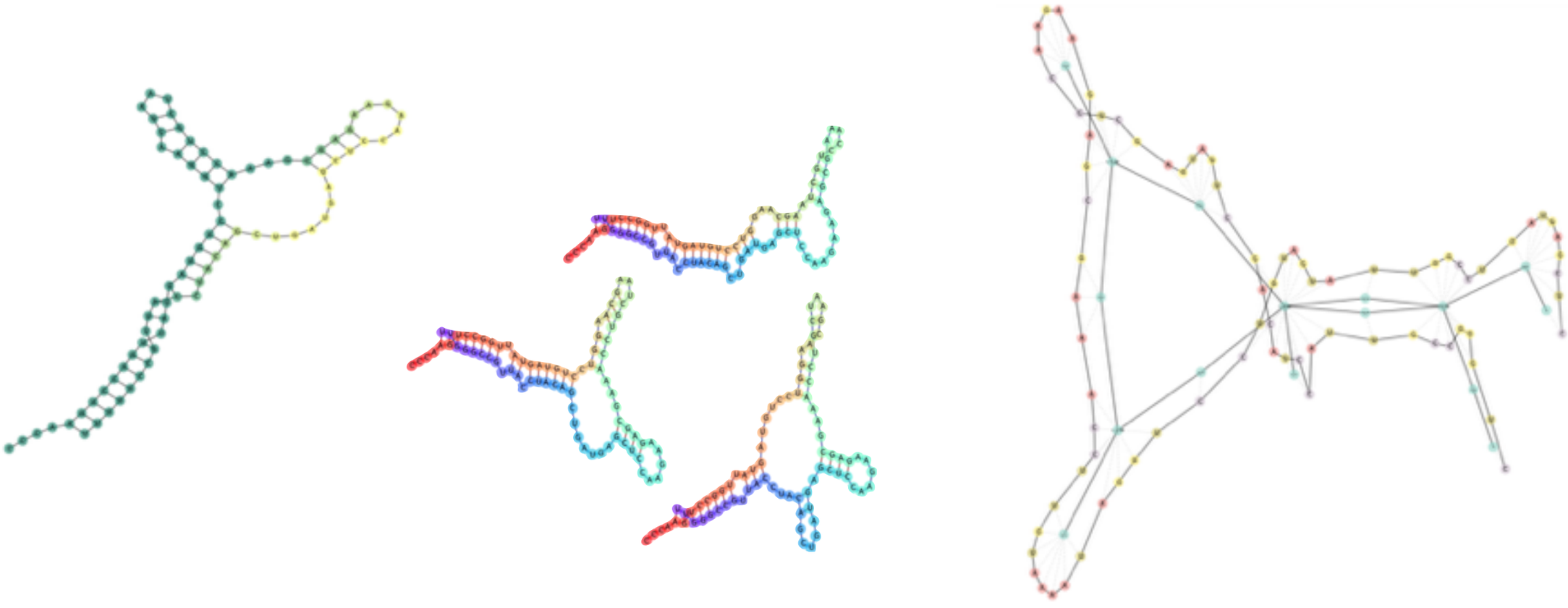
associate, compose, construct



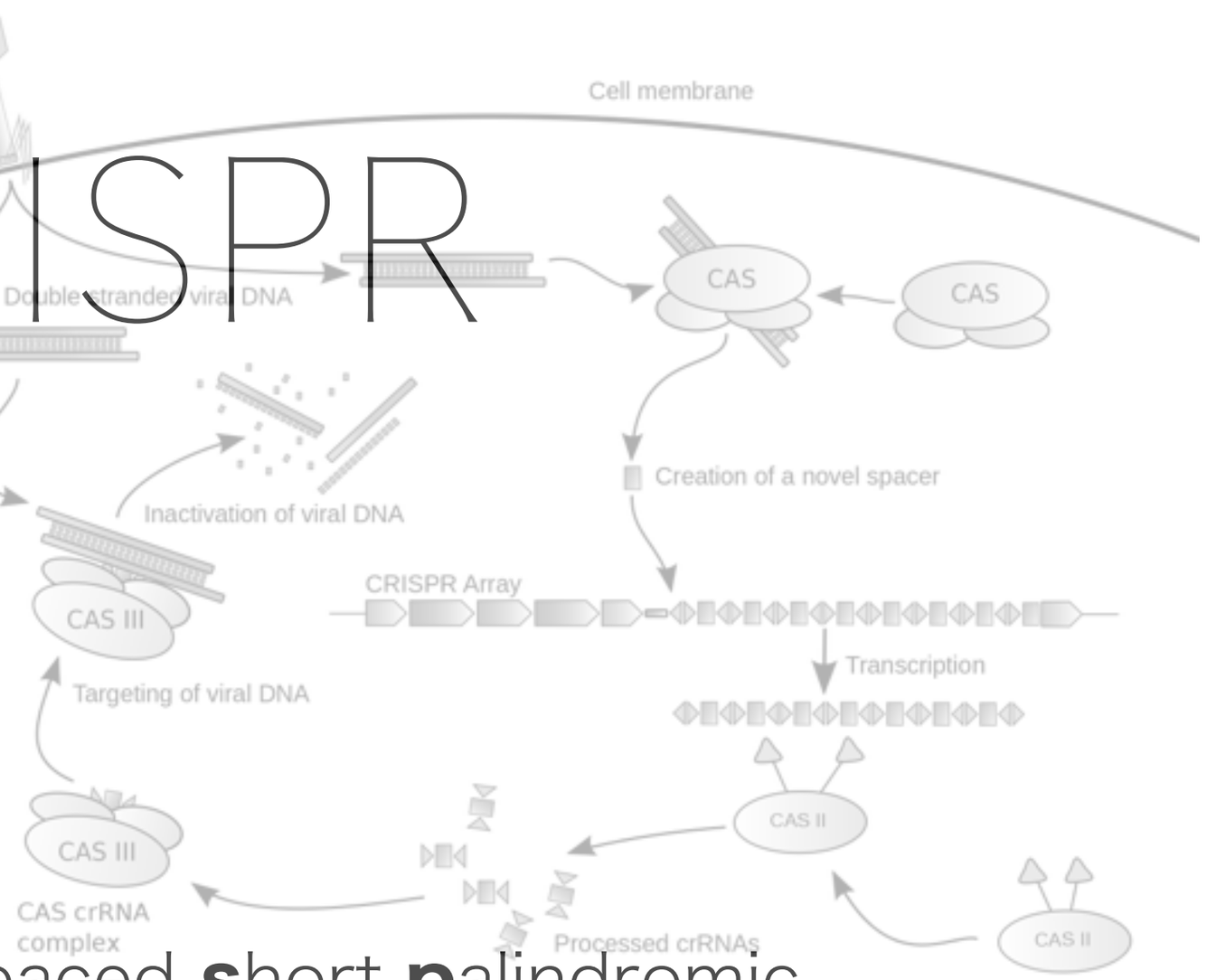




Associate



CRISPR



Clustered **r**egularly-**i**nterspaced **s**hort **p**alindromic **r**epeats are segments of DNA containing short repetitions followed by short segments of DNA from virus or plasmid

The CRISPR/Cas system is a prokaryotic immune system and provides a form of acquired immunity

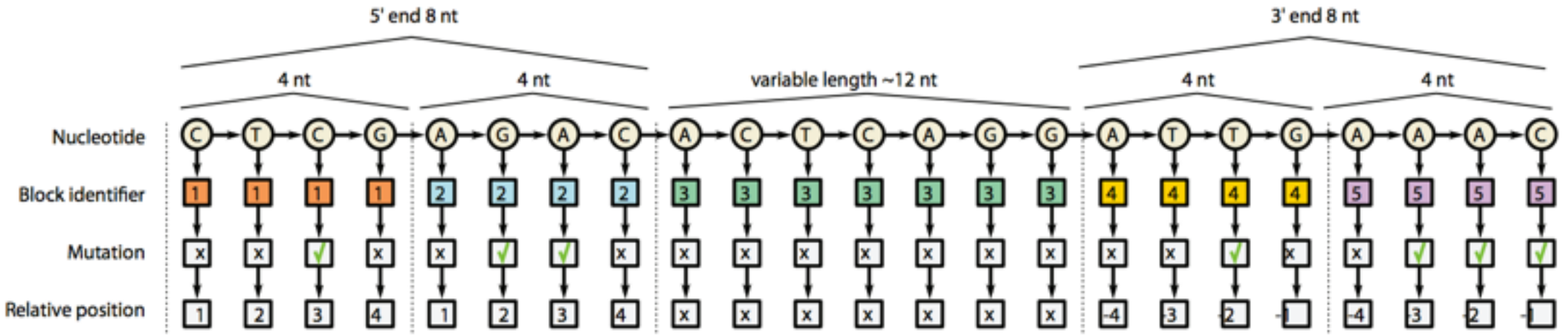
CRISPR-strand

Alkhnbashi, O. S., F Costa, S A. Shah, R A. Garrett, S J. Saunders, R Backofen. Bioinformatics 2014

Prediction of the correct orientation (the strand) of the repeat array is of interest to better define a family notion

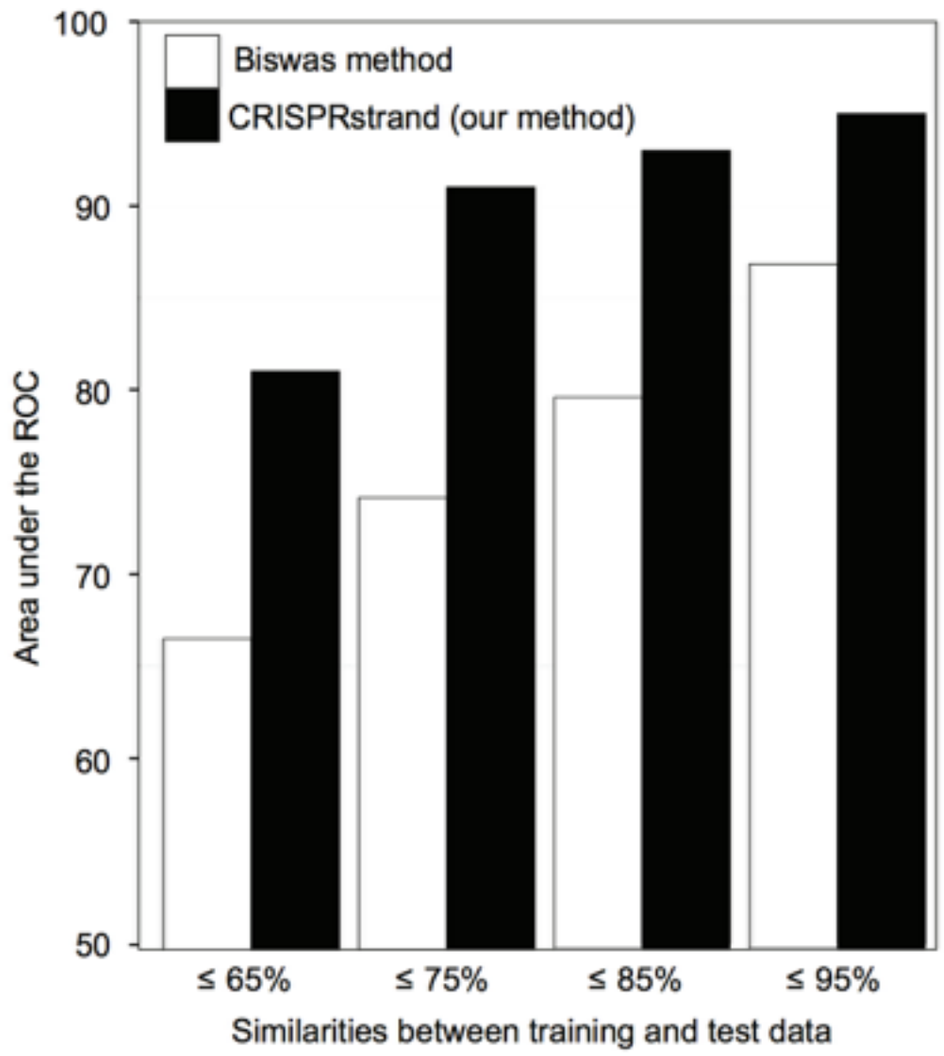
Questions: which parts/characteristics are important > yield better predictive performance

1. k initial nucleotides
2. mutation events
3. relative nucleotides positions and identity



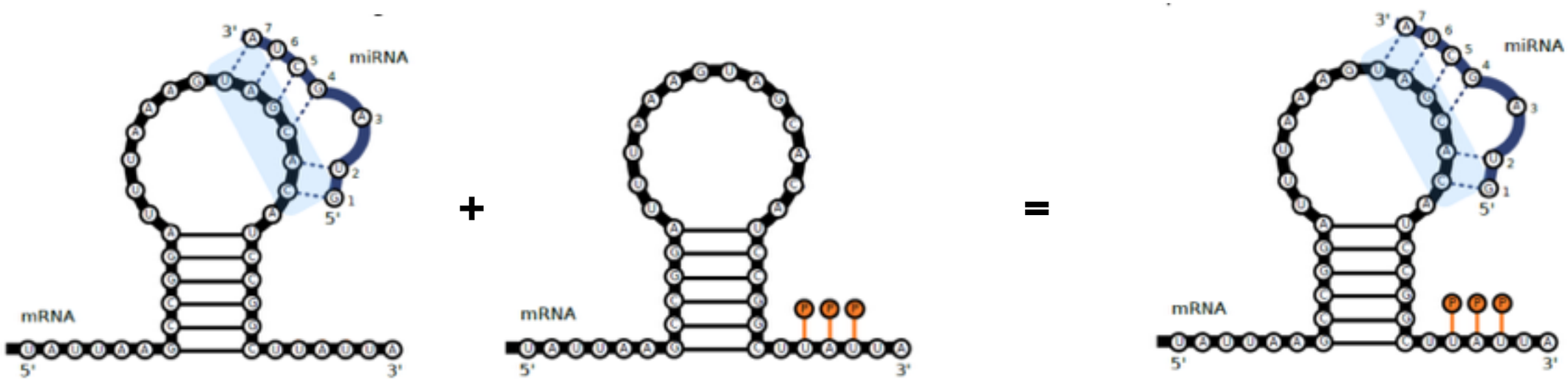
CRISPR-strand

Alkhnbashi, O. S., F Costa, S A. Shah, R A. Garrett, S J. Saunders, R Backofen. Bioinformatics 2014



comparison with traditional method with few hand-crafted features

Compose

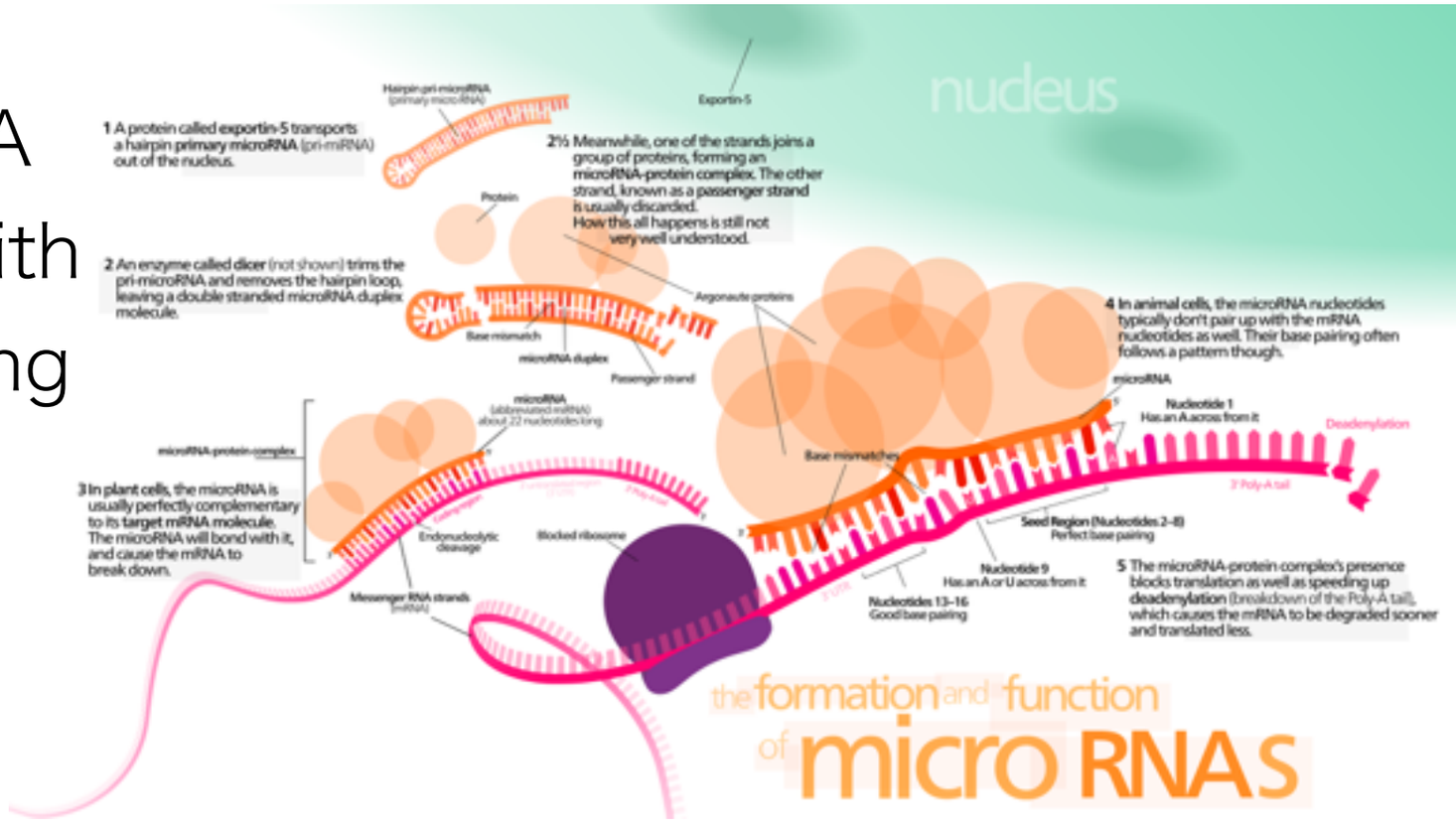


miRNA

- miRNA-RNA interaction

  micro RNA (abbreviated **miRNA**) is a ~22 nucleotides non-coding RNA molecule which regulates post-transcriptionally gene expression

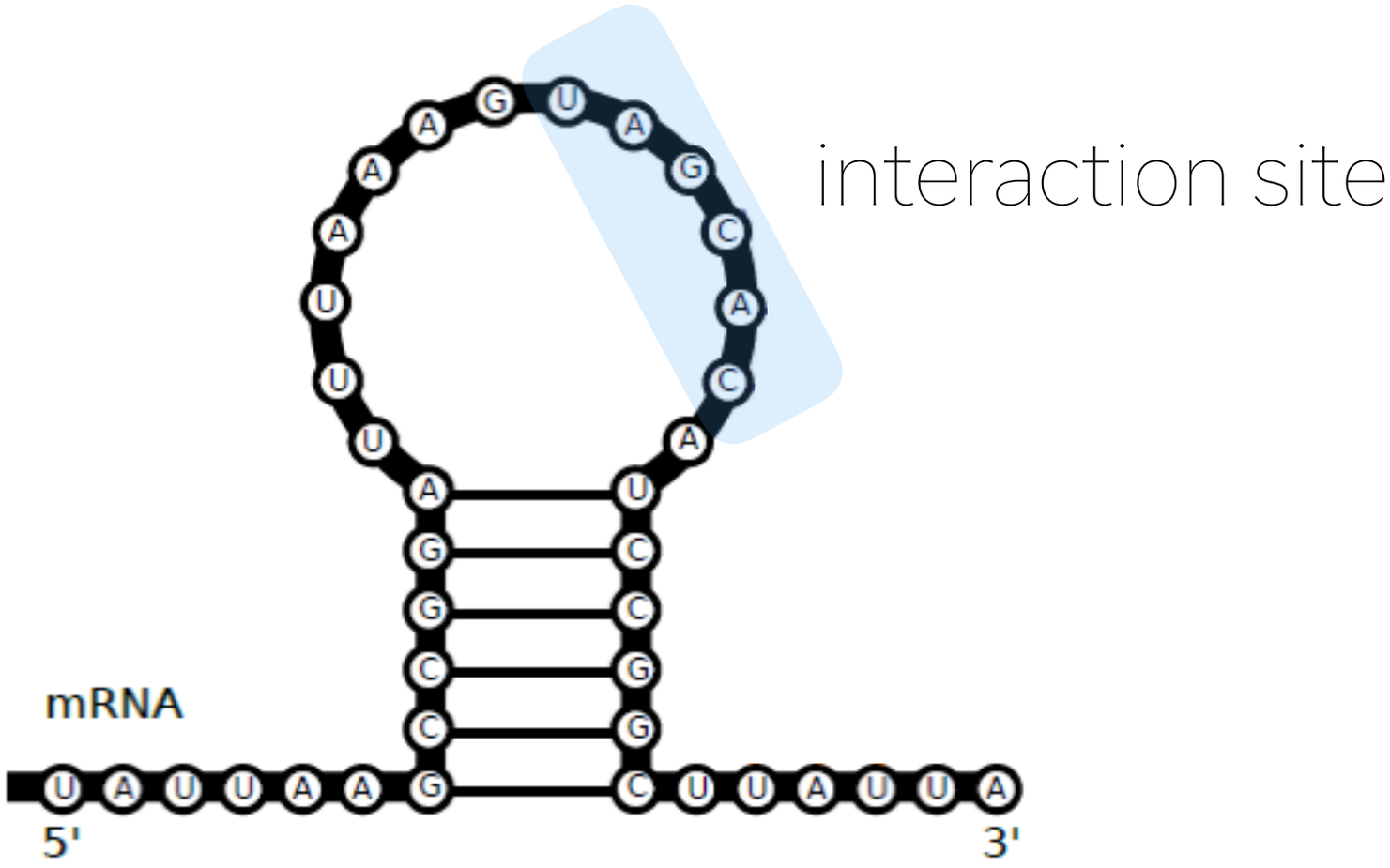
- dysregulation of miRNA has been associated with many diseases including cancer (oncomirs)



mRNA structure

interaction depends on accessibility of mRNA

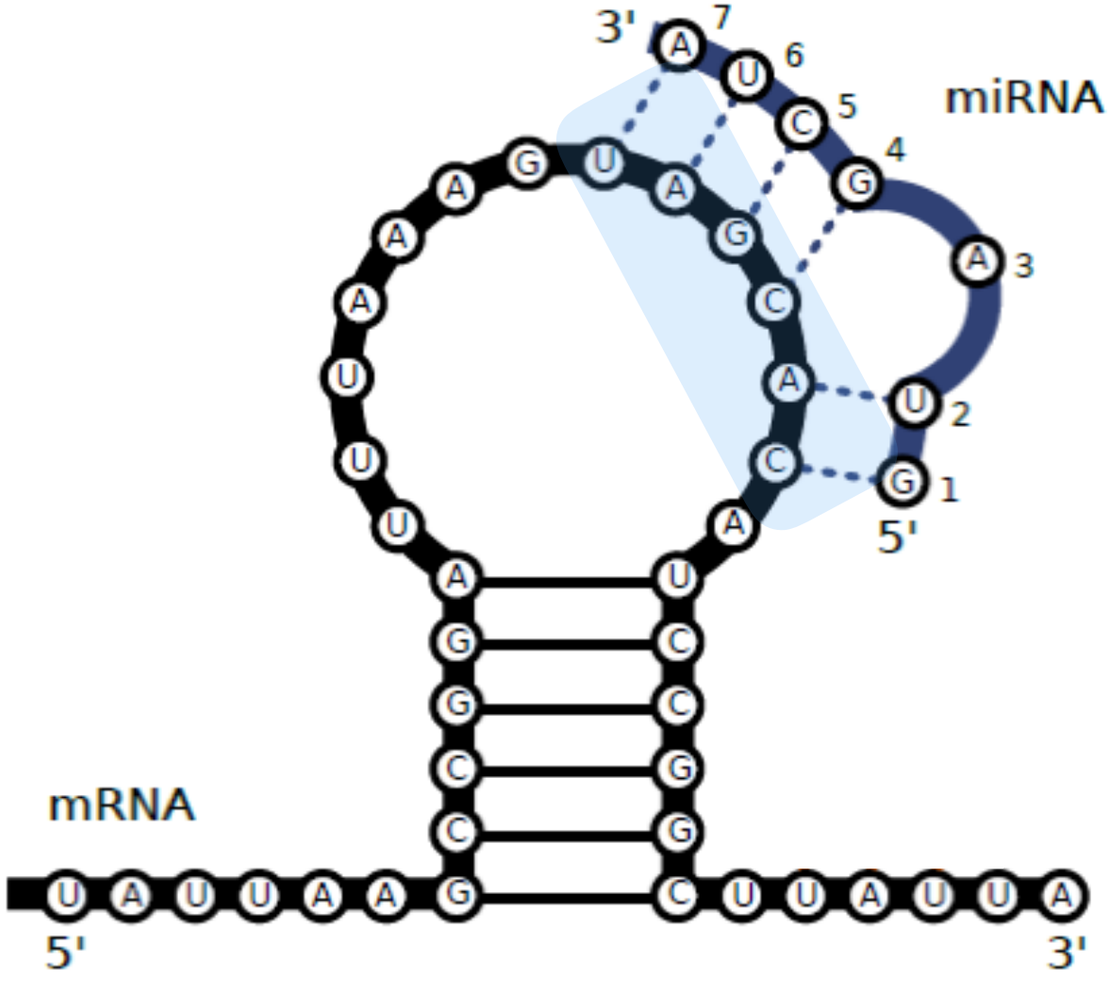
transform mRNA sequence to extract self interacting structure



miRNA-mRNA duplex

interaction depends on duplex stability with miRNA

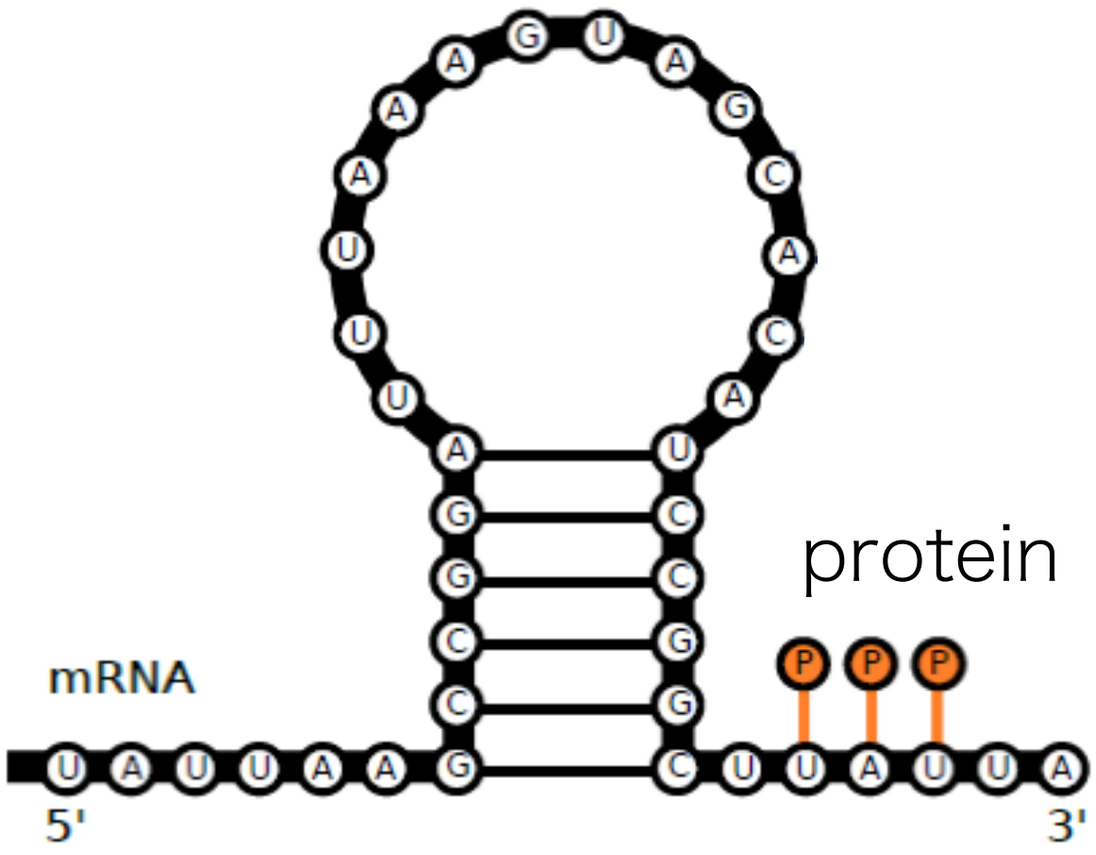
compose folded mRNA with interacting miRNA



mRNA-RBP

interaction depends on presence of RNA binding proteins (RBP) that interact in the neighborhood

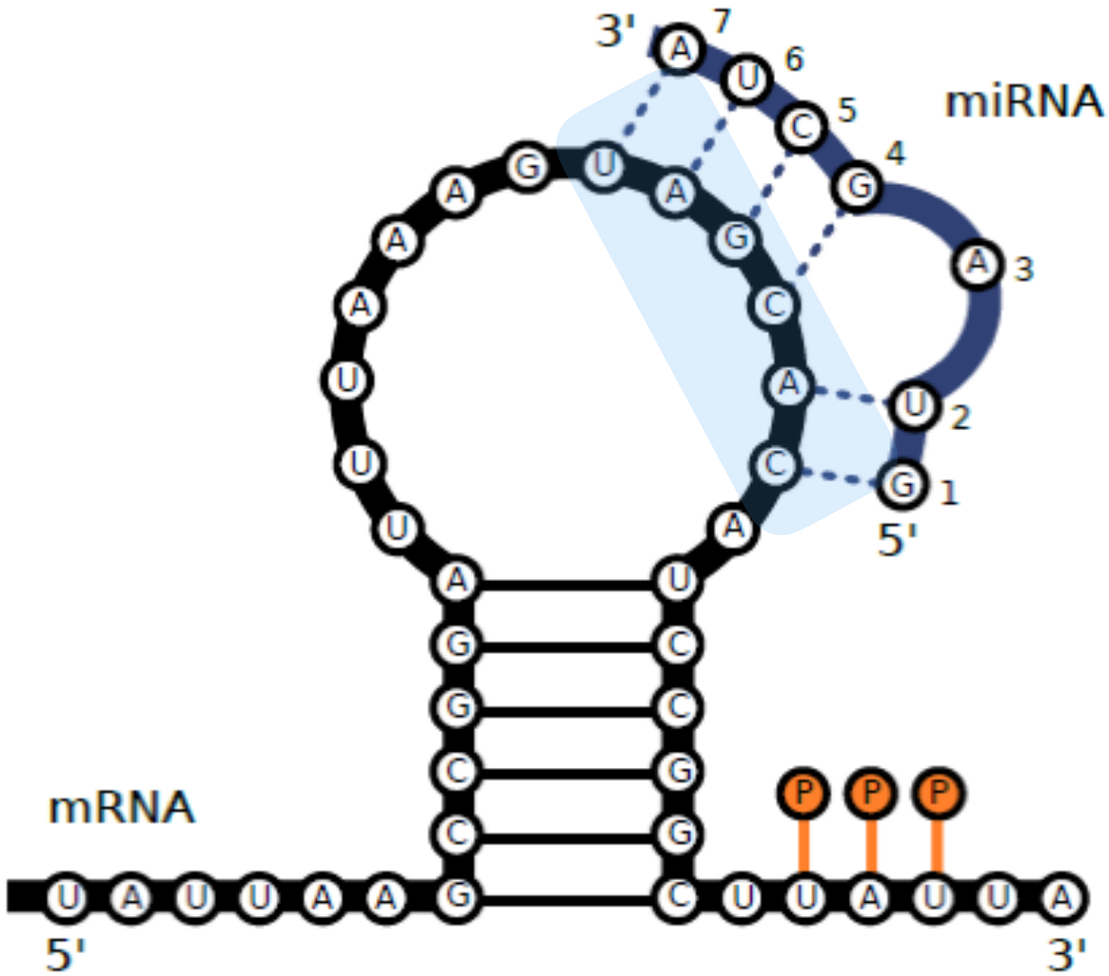
transform folded mRNA with RBP predictor



Full composite

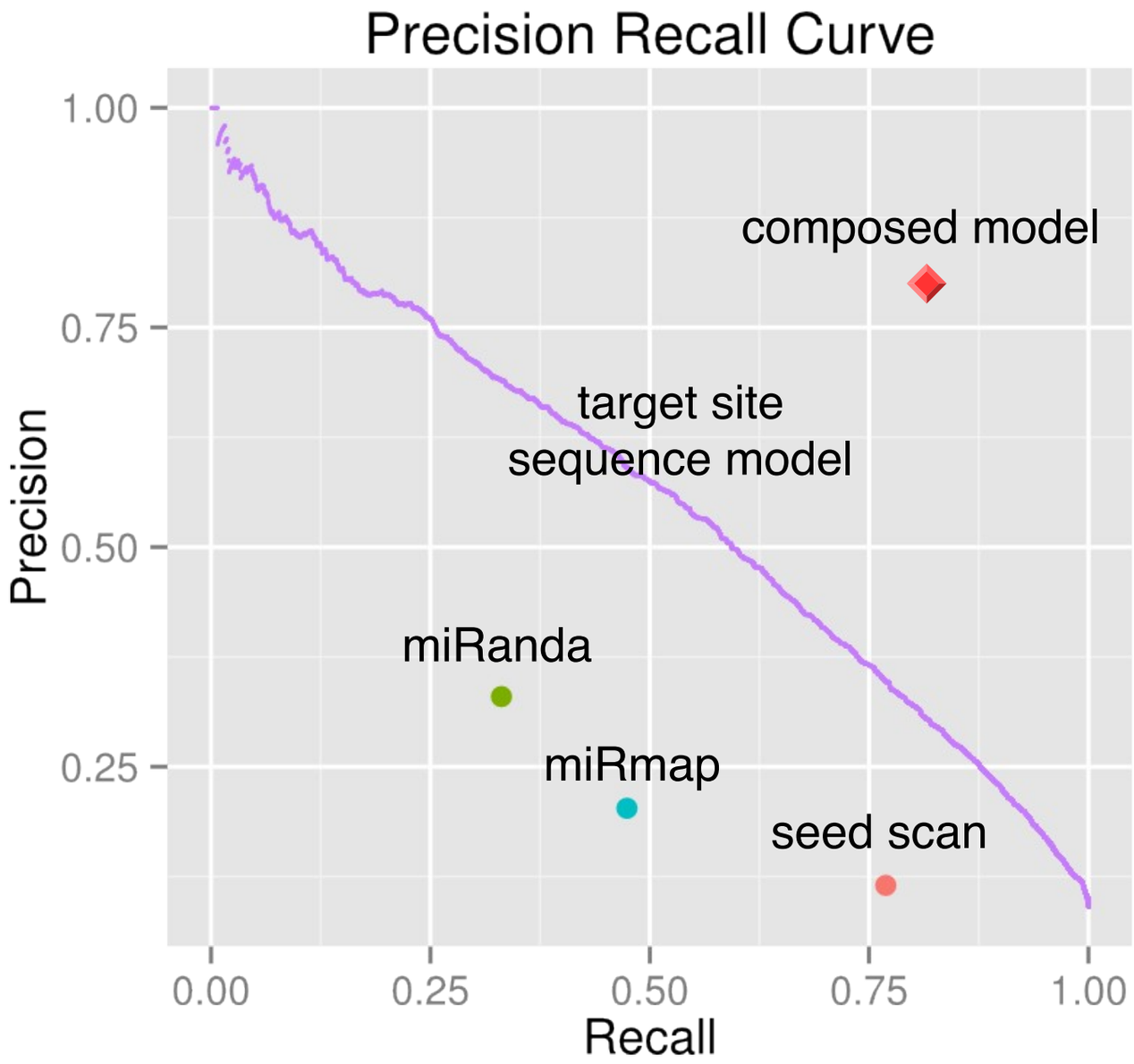
finally since the interaction depends on all factors

compose duplex with RBP interaction

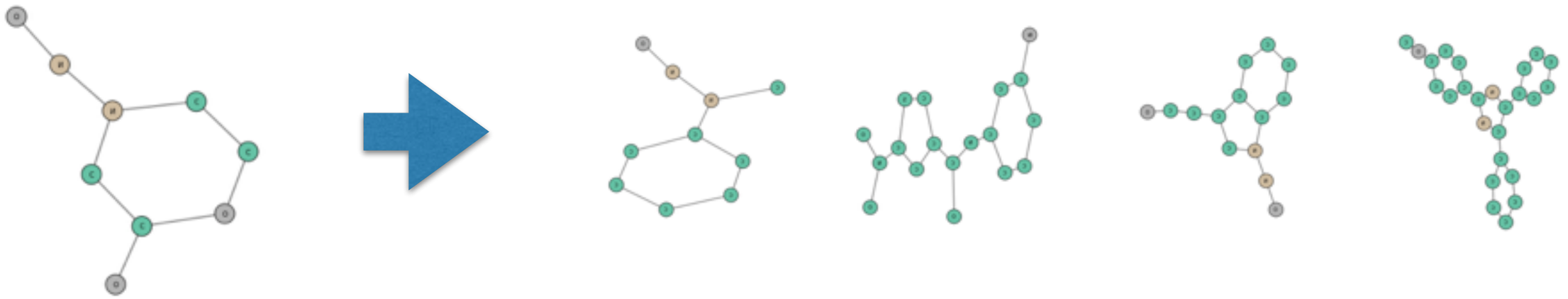


miRNA-mRNA prediction

M. Uhl, F Costa, S J. Saunders, R Backofen (in preparation)



Construct



```
>ABQF01059171.1/305-384
UUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUGCUAAGCAAGGUCC
UGUAGUAUUGGCCUGAACCC
>AADN03003451.1/4511-4593
CUGGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUGUAAAAUAGGUCC
UGUAGUAUUGGCCUGAUGAGCUC
```

```
>A1
UUAGGCCGUUACCUACAGCUGAUGAGCUCCAAGAAGAGCGAAACCUUUUAAGAUAGGUCC
UGUAGUAUUGGCCUGAAAACCAU
>A 2
CUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUGCUAGGUCCUGCAG
UACUGGCUUAAGAGGCUA
>A 3
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUAAGAGAAU
>A 4
UUGAGCCGUUACCUGCAGCUGAUGAGCUCCAAAAAGAGCGAAACCUAUUAGGUCCUGCAG
UACUGGCUUGAGAUA
```

RBP

- RNA binding proteins are proteins that bind to the double or single stranded RNA in cells and participate in forming ribonucleoprotein complexes

- RBPs have crucial roles e.g. cellular function, transport and localization

RBP binding validation

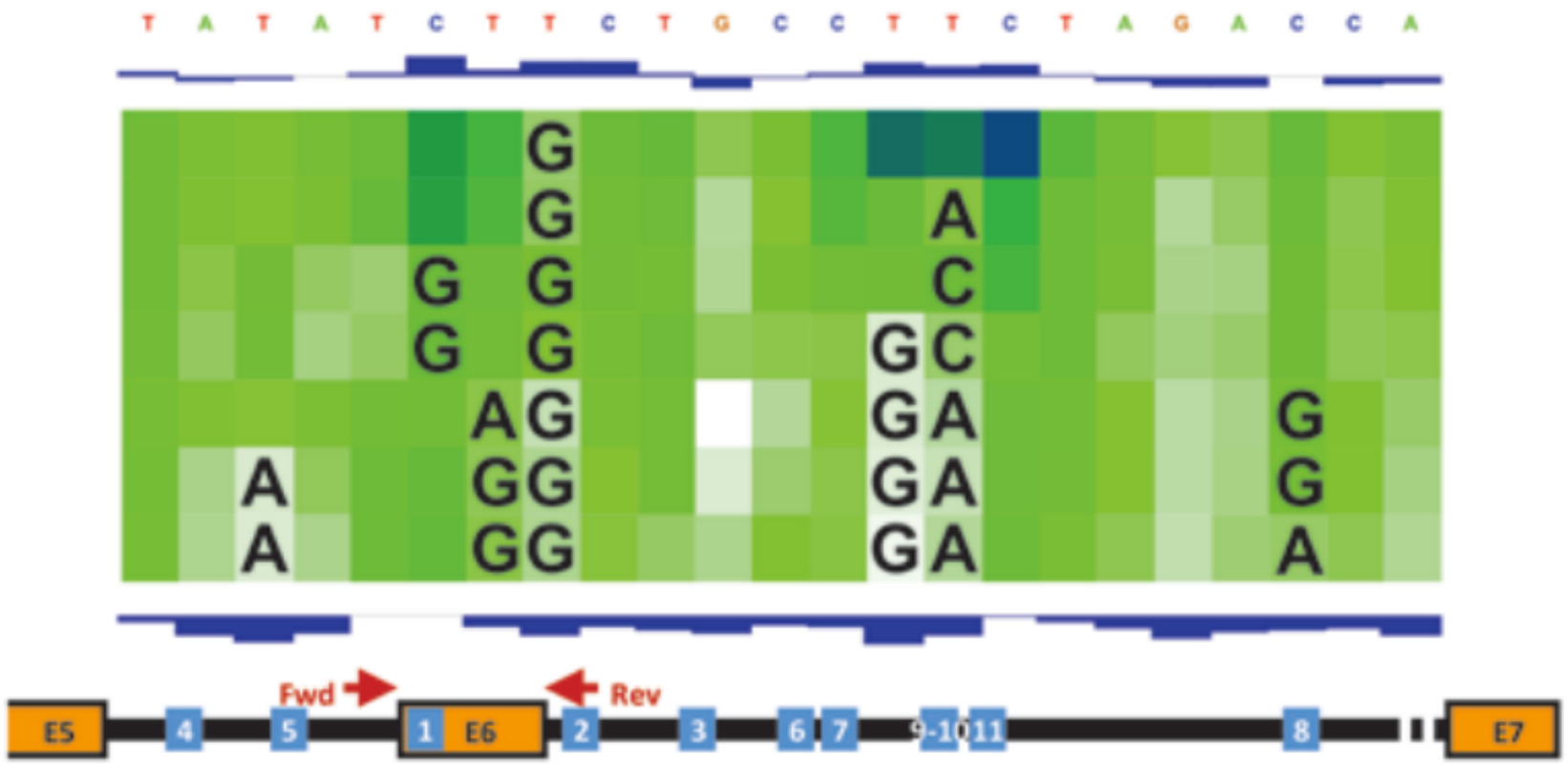
Ferrarese R, Harsh GR, Yadav AK, Bug E, Maticzka D, Reichardt W, Dombrowski SM, Miller TE, Masilamani AP, Dai F, Kim H, Hadler M, Scholtens DM, Yu IL, Beck J, Srinivasasainagendra V, Costa F, Baxan N, Pfeifer D, Elverfeldt DV, Backofen R, Weyerbrock A, Duarte CW, He X, Prinz M, Chandler JP, Vogel H, Chakravarti A, Rich JN, Carro MS, Bredel M, R. et al. J. Clin. Invest. 2014

- excess of PTB protein inhibits splicing ▷ glioblastoma (brain cancer) not repressed efficiently
- identified splicing region with predicted but no experimental evidence for PTB binding: how to validate these sites?
- cannot knock out/down PTB as it mediates many pathways and would result in cell death

RBP binding validation

Ferrarese, R. et al. J. Clin. Invest. 2014

- **idea:** inhibit PTB interaction in specific sites
- **how:** construct alternative sequences that inhibit PTB binding ▷ do prescribed mutations and look for reduction of aberrant splicing



Summing up

- Attempt to develop a vocabulary to describe complex analytical processes at an useful abstract level as functional composition. E.g.

associate[a](compose[b]([transform[c](convert(X)), transform[d](X)]))

≈ SVM(RNAFold(X)+ RBP(X))

- EDeN library provides support for efficient implementation of graph to vector mapping ▷use of state-of-the-art ML libraries

- GArDen (Generic Abstract Decomposition) library **will** provide support for the generic computational framework and support for
  - automatic parallelization and
  - hyper parameter optimization

Conclusion

- A new kind of (life) science is appearing:
  - not only study of nature, but simultaneous engineering using available partial knowledge
    - with a growing need for computational tools to:
      1. make sense of BIG and heterogeneous data
      2. support causal relationships investigation
      3. support rational synthesis design

- There is a need to upgrade analytical systems for the synthetic scientific era

Thank you

Acknowledgments:

Frasconi P, De Raedt L, Backofen R, De Grave K, Schietgat L, Ramon J, Alkhnbashi OS, Shah SA, Garrett RA, Saunders SJ, Ceroni A, Verbeke M, Kundu K, Huber M, Reth M, Mann M, Makarova KS, Wolf YI, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, van der Oost J, Koonin EV, Videm P, Rose D, Menchetti S, Ferrarese R, Harsh GR, Yadav AK, Bug E, Maticzka D, Reichardt W, Dombrowski SM, Miller TE, Masilamani AP, Dai F, Kim H, Hadler M, Scholtens DM, Yu IL, Beck J, Srinivasasainagendra V, Baxan N, Pfeifer D, Elverfeldt DV, Weyerbrock A, Duarte CW, He X, Prinz M, Chandler JP, Vogel H, Chakravarti A, Rich JN, Carro MS, Bredel M, Passerini A, Pollastri G, Pontil M, Corrado G, Tebaldi T, Bertamini G, Quattrone A, Viero G,