Three Grand Challenges in Computational Biology

Tandy Warnow
Founder Professor of Engineering
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Phylogenomics

From the Tree of the Life Website,
University of Arizona
Metagenomics

• Analyses of environmental samples, containing mixtures of organisms.

• Typical metagenomics dataset: millions of reads generated by some next generation sequencing machine (e.g., Illumina).
Multiple Sequence Alignment (MSA): 
a scientific grand challenge

Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

---

Frontiers in Massive Data Analysis, National Academies Press, 2013
Three Grand Challenges

- Multiple sequence alignment
- Phylogeny estimation and the Tree of Life
- Metagenomic taxon identification
Today’s talk: methods for

• Multiple sequence alignment
• Phylogeny estimation and the Tree of Life
• Metagenomic taxon identification
  • A unifying technique is the “Ensemble of Hidden Markov Models” (introduced by Mirarab et al., 2012)
The “Tree of Life”

Basic Biology:
• How did life evolve?
• What are the mechanisms of molecular evolution?
• How do species adapt to environmental changes?

Applications of phylogenies to:
protein structure and function
population genetics
human migrations
metagenomics
**phylogenomics**

Two dimensions: number of genes and number of species
Avian Phylogenomics Project

- E Jarvis, HHMI
- MTP Gilbert, Copenhagen
- G Zhang, BGI
- T. Warnow, UT-Austin
- S. Mirarab, UT-Austin
- Md. S. Bayzid, UT-Austin

Plus many many other people...

- Approx. 50 species, whole genomes, 14,000 loci
- Jarvis, Mirarab, et al., Science 2014

Major challenges:
- 200+ CPU years to perform “concatenated maximum likelihood analysis”
- Massive gene tree heterogeneity consistent with incomplete lineage sorting
1kp: Thousand Transcriptome Project

First paper Wickett, Mirarab, et al., *PNAS* 2014

Second study will be much bigger: Plant Tree of Life based on transcriptomes of ~1200 species and more than 13,000 gene families (most not single copy)

Upcoming challenges:
- Massive gene tree conflict consistent with ILS
- Alignments and gene trees on >100,000 sequences
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology
Phylogenomic pipeline

• Select taxon set and markers
• Gather and screen sequence data, possibly identify orthologs
• Compute multiple sequence alignments for each locus, and construct gene trees
• **Compute species tree or network:**
  – Combine the estimated gene trees, OR
  – Estimate a tree from a concatenation of the multiple sequence alignments
• Get statistical support on each branch (e.g., bootstrapping)
• Estimate dates on the nodes of the phylogeny
• Use species tree with branch support and dates to understand biology
Phylogenomic pipeline

• Select taxon set and markers
• Gather and screen sequence data, possibly identify orthologs
• Compute multiple sequence alignments for each locus, and construct gene trees
• Compute species tree or network:
  – Combine the estimated gene trees, OR
  – Estimate a tree from a concatenation of the multiple sequence alignments
• Get statistical support on each branch (e.g., bootstrapping)
• Estimate dates on the nodes of the phylogeny
• Use species tree with branch support and dates to understand biology
Sequence Evolution (simplified)
Indels (insertions and deletions)
The true multiple alignment
- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Phase 1: Alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
Simulation Studies

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCACG
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCACG
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

Unaligned Sequences

Compare

True tree and alignment

Estimated tree and alignment
Quantifying Error

**TRUE TREE**

**DNA SEQUENCES**
- $S_1$: ACAATTAGAAC
- $S_2$: ACCCTTAGAAC
- $S_3$: ACCATTCCAAC
- $S_4$: ACCAGACCAAC
- $S_5$: ACCAGACCAGGA

**INFERRED TREE**

**FN**: false negative (missing edge)
**FP**: false positive (incorrect edge)

50% error rate
1000-taxon models, ordered by difficulty (Liu et al., 2009)
Large-scale Alignment Estimation

• Many genes are considered unalignable due to high rates of evolution

• Only a few methods can analyze large datasets

• Phylogenomics projects need better methods!
Re-aligning on a tree

Decompose dataset

Estimate ML tree on merged alignment

Align subsets

Merge sub-alignments
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment

Alignment

Tree

Repeat until termination condition, and return the alignment/tree pair with the best ML score
SATé-1 (Science 2009) performance

1000-taxon models, ordered by difficulty –
rate of evolution generally increases from left to right

SATé-1 24 hour analysis, on desktop machines (similar improvements for biological data)

SATé-1 can analyze up to about 8,000 sequences.
1000-taxon models ranked by difficulty

SATé-1: up to 8K and SATé-2: up to ~50K
PASTA
Tree accuracy

1 million sequences:

- PASTA finished one iteration in 15 days
- PASTA tree had 6% error, compared to 5.6% when using true alignment
- Starting tree had 8.4% error

PASTA: Mirarab et al., RECOMB 2014 and J. Comp Biol 2015
PASTA and SATé: meta-methods

Decompose dataset → Align subsets

Estimate ML tree on merged alignment

Merge sub-alignments
1kp: Thousand Transcriptome Project

- First paper Wickett, Mirarab, et al., *PNAS* 2014
- Second study will be much bigger: Plant Tree of Life based on transcriptomes of ~1200 species and more than 13,000 gene families (most not single copy)

Upcoming Challenges:
- Massive gene tree conflict consistent with ILS
- **Alignments and gene trees on >100,000 sequences**

Plus many many other people…
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary

All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences with many fragmentary sequences
UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”


Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.
UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”


Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.

Uses an ensemble of HMMs
Simple idea (not UPP)

• Select random subset of sequences, and build “backbone alignment”

• Construct a Hidden Markov Model (HMM) on the backbone alignment

• Add all remaining sequences to the backbone alignment using the HMM
This approach works well if the dataset is small and has low evolutionary rates, but is not very accurate otherwise.

- Select random subset of sequences, and build “backbone alignment”
- Construct a Hidden Markov Model (HMM) on the backbone alignment
- Add all remaining sequences to the backbone alignment using the HMM
One Hidden Markov Model for the backbone alignment?
Or 2 HMMs?
Or 4 HMMs?
Or all 7 HMMs?
UPP Algorithmic Approach

1. Select random subset of full-length sequences, and build “backbone alignment”

2. Construct an “Ensemble of Hidden Markov Models” on the backbone alignment

3. Add all remaining sequences to the backbone alignment using the Ensemble of HMMs
Evaluation

• Simulated datasets (some have fragmentary sequences):
  – 10K to 1,000,000 sequences in RNASim – complex RNA sequence evolution simulation
  – 1000-sequence nucleotide datasets from SATé papers
  – 5000-sequence AA datasets (from FastTree paper)
  – 10,000-sequence Indelible nucleotide simulation

• Biological datasets:
  – Proteins: largest BaliBASE and HomFam
  – RNA: 3 CRW datasets up to 28,000 sequences
Notes:

• We show alignment error using average of SP-FN and SP-FP.

• UPP variants have better alignment scores than PASTA.

• (Not shown: Total Column Scores – PASTA more accurate than UPP)

• No other methods tested could complete on these data
RNASim Million Sequences: tree error

Using 12 processors:

- UPP(Fast,NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days
UPP is very robust to fragmentary sequences

Under high rates of evolution, PASTA is badly impacted by fragmentary sequences (the same is true for other methods).

UPP continues to have good accuracy even on datasets with many fragments under all rates of evolution.

Performance on fragmentary datasets of the 1000M2 model condition
UPP Running Time

Wall-clock time used (in hours) given 12 processors
Summary so far

Most MSA methods degrade in accuracy with evolutionary distances, dataset size, and/or fragmentation.

PASTA and SATé use divide-and-conquer and iteration to improve accuracy and scalability of base MSA methods. They have excellent accuracy, but (like their base methods) are impacted by fragmentation.

By design, HMMs are less impacted by fragmentation, but single HMMs are not as accurate as ensembles of HMMs.

UPP uses an Ensemble of HMMs to improve accuracy compared to a single HMM, and is much more accurate than all other methods tested in the presence of fragments.
Most MSA methods degrade in accuracy with evolutionary distances, dataset size, and/or fragmentation.

PASTA and SATé use divide-and-conquer and iteration to improve accuracy and scalability of base MSA methods. They have excellent accuracy, but (like their base methods) are impacted by fragmentation.

By design, HMMs are less impacted by fragmentation, but single HMMs are not as accurate as ensembles of HMMs.

UPP uses an Ensemble of HMMs to improve accuracy compared to a single HMM, and is much more accurate than all other methods tested in the presence of fragments.
Metagenomics

- Analyses of environmental samples, containing mixtures of organisms.

- Typical metagenomics dataset: millions of reads generated by some next generation sequencing machine (e.g., Illumina).
The NIH Human Microbiome Project

25,000 human genes, 1,000,000 bacterial genes
Discovery of new genes and species

Metagenomics, Venter et al., Exploring the Sargasso Sea: Scientists Discover One Million New Genes in Ocean Microbes
Basic Questions

What is this fragment? (Classify each fragment as well as possible.)

What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

What are the organisms in this metagenomic sample doing together?
Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample
Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

For example: The distribution of the sample at the species-level is:

- 50% species A
- 20% species B
- 15% species C
- 14% species D
- 1% species E
TIPP: Taxonomic Identification using Phylogenetic Profiles

TIPP: taxon identification and phylogenetic profiling (Nguyen et al., Bioinformatics, 2014)
TIPP pipeline

Input: set of reads from a shotgun sequencing experiment of a metagenomic sample.

1. Assign reads to marker genes using BLAST
2. For reads assigned to marker genes, perform taxonomic analysis
3. Combine analyses from Step 2
TIPP pipeline

Input: set of reads from a shotgun sequencing experiment of a metagenomic sample.

1. Assign reads to marker genes using BLAST
2. For reads assigned to marker genes, perform taxonomic analysis
3. Combine analyses from Step 2
Phylogenetic Placement

Fragmentary sequences from some gene

Full-length sequences for same gene, with an alignment and a tree
Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

Step 2: Place each query sequence into backbone tree, using extended alignment
Align Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = TAAAAC
Align Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = --------T-A--AAAC--------
Place Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-------GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = ---------T-A--AAAC--------
Phylogenetic Placement

• Build an HMM for the backbone alignment, and align each query sequence to the backbone alignment using the HMM.

• Place each query sequence into backbone tree optimizing maximum likelihood.
Phylogenetic Placement

• Build an HMM for the backbone alignment, and align each query sequence to the backbone alignment using the HMM.

• Place each query sequence into backbone tree optimizing maximum likelihood.
TIPP

Technique:
• Aligns query sequences using a disjoint ensemble of HMMs
• Finds maximum likelihood placement in reference tree
• Considers statistical support for each alignment and phylogenetic placement

Note: can control precision/recall tradeoff by modifying support thresholds.

Results: improves on the use of a single HMM for both taxon identification and abundance profiling.
Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

Leading techniques:

- **PhymmBL** (Brady & Salzberg, Nature Methods 2009)
- **NBC** (Rosen, Reichenberger, and Rosenfeld, Bioinformatics 2011)
- **MetaPhyler** (Liu et al., BMC Genomics 2011), from the Pop lab at the University of Maryland
- **MetaPhlAn** (Segata et al., Nature Methods 2012), from the Huttenhower Lab at Harvard
- **mOTU** (Bork et al., Nature Methods 2013)

MetaPhyler, MetaPhlAn, and mOTU are marker-based techniques (but use different marker genes).

Marker gene are single-copy, universal, and resistant to horizontal transmission.
High indel datasets containing known genomes

Note: NBC, MetaPhlAn, and MetaPhyler cannot classify any sequences from at least one of the high indel long sequence datasets, and mOTU terminates with an error message on all the high indel datasets.
"Novel" genome datasets

Note: mOTU terminates with an error message on the long fragment datasets and high indel datasets.
TIPP compared to other abundance profilers

• TIPP is highly accurate, even in the presence of high indel rates and novel genomes, and for both short and long reads.

• All other methods have some vulnerability (e.g., mOTU is only accurate for short reads and is impacted by high indel rates).

Future sequencing technologies likely to have higher indel rates and produce longer reads!
Summary

• **SATé-1** (Science 2009), **SATé-2** (Systematic Biology 2012), co-estimation of alignments and trees. SATé-2 is well established in the biology community (e.g., used in Jarvis, Mirarab et al., Science 2014 avian phylogenomics project).

• **PASTA** (RECOMB 2014 and J Comp Biol 2015) is the replacement for SATé. PASTA can analyze up to 1,000,000 sequences.

• **UPP** (ultra-large multiple sequence alignment), Genome Biology 2015. UPP produces highly accurate alignments, even in the presence of fragmentary sequences and can analyze datasets with 1,000,000 sequences.

• **TIPP** (metagenomic taxon identification and abundance profiling), Bioinformatics 2014, is highly accurate and more robust to error rates and variation in read length than other methods. TIPP is also able to perform well on novel sequences, where other methods often fail.
Other Applications of the Ensemble of HMMs

HIP-HOP (protein classification and remote homology detection), submitted

Protein structure and function prediction

Mass spec

Metagenomic taxonomic identification and abundance profiling, and metagenome assembly
This meant that a single cross-fold set of test data required over 2,500 node hours, whereas all other methods ran in less than 24 node hours. As a result, the data shown here for all methods only include results on a single cross-fold set, which includes 312,841 query sequences of each fragment length (100% full-length, 50% length, and 25% length; 938,523 sequences total). On all other methods the precision and recall results were virtually identical across all four cross-fold sets (see SI Exhibit 1 for details).

Figure 3 contains precision-recall curves for each of the four different methods under consideration. The curves are estimated by varying an inclusion threshold parameter for the particular method and producing five to seven distinct points, with intermediate values interpolated linearly. A table with all values for all points in Figure 3 is available in the Appendix A2.

The precision-recall results show that for all sequence lengths, HIP-HOP dominates all other methods at every one of its computed points on the curve. More specifically, HIP-HOP appears to be doing well by way of highly improved recall without a strong trade-off in precision. In particular on full-length sequences, the increase in recall is roughly 0.5% for the same precision, which is modest but statistically significant with \( p < .001 \) using a binomial test. However, on the fragmentary sequences HIP-HOP provides a substantial increase in recall compared to HMMER and BLAST. For example, at a 99.2% precision, on 50% length sequences HIP-HOP’s recall was 6% greater than HMMER and 16% greater than BLAST, and on 25% length sequences it is 17% greater than HMMER and 10% greater than BLAST.

Figure 7 contains additional precision-recall curves that are similar to the previous figure but families are grouped according to the size of the seed alignment and its average pairwise sequence identity. The grouping by size has only two levels: families with 0 to 100 seed sequences and those with more than 100 sequences. The grouping by average pairwise sequence identity has three levels: 0-20%, 20-30%, and more than 30%, which represent about 5%, 28% and 67% of the 11,156 families, respectively.

The amount of sequence identity has been known to be related to the difficulty of remote homology detection; in particular Rost (1999) identified the 20-35% sequence identity level as a rough point below which homology detection is substantially more challenging. Moreover, previous research has shown that a single HMM on a large and evolutionarily divergent data set can lead to poor downstream analyses, and by using an ensemble of HMMs (as in the SEPP, TIPP, and UPP studies) the accuracy of the analyses can be boosted. These two dimensions are therefore of particular interest to examine the performance of our methods at the subgroup level.

It is clear from Figure 7 that HIP-HOP’s improvement over the other methods is not localized to one part of the data space; the HIP-HOP precision-recall curve is the most outward from the origin in nearly every case. Nonetheless, the effect appears to be particularly strong on the low sequence-identity families (those with <20% and 20-30% identity). In the center and right columns, the distance between HIP-HOP (in red) and HMMER (blue) is the most consistent and the largest. Additionally, the benefits of the HIP-HOP method in terms of both precision and recall are more pronounced on the smaller data sets.

Finally, one additional observation from Figure 7 is interesting. BLAST substantially outperforms a single HMM on the four “hardest” subgroups in this analysis, the quarti-length fragments on families with <20% and 20-30% pairwise identity, and in the two hardest cases has equal or better performance to HIP-HOP. While BLAST and HMMER use fundamentally different techniques scoring the similarity between a pair of sequences or a sequence and an HMM, in an abstract sense, the strategy employed by HIP-HOP is essentially a compromise between BLAST, which searches every individual sequence in a family for similarity, and HMMER, which searches only one HMM representing the whole group. HIP-HOP does a more granular search than HMMER, but not to the level of one per target-family sequence. These results suggest that the more evolutionarily divergent a family is, the more a granular search may be necessary, and an interesting course of future research may be to characterize this relationship better.

**5 Conclusion**

The HIP-HOP method uses a technique for improving the use of single prolic HMMs to represent a seed alignment, and to detect membership in protein families and superfamilies. HIP-HOP has a very powerful algorithmic design, and does not need to use or estimate structural features to obtain its improvement in accuracy.

One novel algorithmic advancement presented in this study is the use of the dynamic decomposition to partition the subset sizes according to the empirical statistics of the data (i.e., the mean pairwise sequence identity of the induced alignment). Previous applications of the ensemble of HMMs technique differed in various respects, but their decompositions used stopping rules that only considered either the number of sequences in each subset, or the total number of subsets created. This study shows that additional accuracy can be obtained by considering also the features of the subsets (see Appendix A.1).
Scientific challenges:

- Ultra-large multiple-sequence alignment
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution

Techniques:

- Machine learning
- Applied probability theory
- Graph theory
- Combinatorial optimization
- Supercomputing
- Heuristics
PASTA: Siavash Mirarab (now at UCSD), Nam-phuong Nguyen (now postdoc at UIUC)
UPP: Nam-phuong Nguyen, Siavash Mirarab, and Keerthana Kumar (undergrad)
TIPP: Nam-phuong Nguyen, Siavash Mirarab, Bo Liu, and Mihai Pop

Current NSF grants:

- ABI-1458652 (multiple sequence alignment)
- III:AF:1513629 (metagenomics –with Maryland)
- DBI:1461364 (phylogenomics – with Rice and Stanford)
- CCF:1535977 (graph algorithms to improve phylogenetic estimation - with Berkeley)

Other support: Guggenheim Foundation, Microsoft Research New England, NSF DEB: 0733029, NSF DBI:1062335, David Bruton Jr. Centennial Professorship, the University of Alberta (Canada), Grainger Foundation (at UIUC), UIUC; HHMI (to Siavash Mirarab)

Computational analyses performed on TACC and BlueWaters
HIPHOP: Experimental Design

• Dataset
  • All seed sequences from the PFAM database with at least 10 seed sequences
  • 11,156 PFAM families total

• Methods
  • HIPHOP (Nguyen at al. in submission)
  • HHSearch (Soding 2005)
  • HMMER (Eddy 2011)
  • BLAST (Altschul 1990)

• Experimental pipeline
  • Partitioned the seed sequences into a training set and a query set
  • Scored the assignments of the query sequences to the PFAMs
  • Also generated fragmentary versions of the seed sequences to examine performance on fragmentary data
HIPHOP: Precision-Recall Curves

```
<table>
<thead>
<tr>
<th>Method</th>
<th>Seq Length: Full</th>
<th>Seq Length: 50%</th>
<th>Seq Length: 25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST</td>
<td>Precision: .98</td>
<td>Precision: .92</td>
<td>Precision: .90</td>
</tr>
<tr>
<td>HHSearch</td>
<td>Precision: .98</td>
<td>Precision: .95</td>
<td>Precision: .80</td>
</tr>
<tr>
<td>HIP-HOP</td>
<td>Precision: .98</td>
<td>Precision: .98</td>
<td>Precision: .90</td>
</tr>
<tr>
<td>HMMER</td>
<td>Precision: .96</td>
<td>Precision: .50</td>
<td>Precision: .50</td>
</tr>
</tbody>
</table>
```
Results across model conditions

Avg. Pairwise Sequence Identity

- > 30%
- 20-30%
- < 20%

Sequence Length: Full
Size: > 100

Sequence Length: 25%
Size: > 100

Sequence Length: 50%
Size: > 100

Recall

Precision
Related Research

- Historical linguistics, 1994-present
- Absolute fast converging methods 1997-2002
- Phylogenetic networks, 2003-2005
- Genome rearrangements, 2000-2006
- Multiple sequence alignment, 2009-present
- Supertree methods, 2009-present
- Metagenomic analysis, 2014-present
- Coalescent-based species tree estimation (2011-present)
- Protein classification and remote homology detection, 2015-present