A Retrospective on Genomic Processing for Comparative Genomics

Binhai Zhu

Computer Science Department

Montana State University

Bozeman, MT

USA

1. David Sankoff's Contribution: My Personal Experience

- First heard David's name in 1994.
- First email contact for COCOON'03.
- First switched to computational biology in 2004/5, one of the first problems I worked on was exactly posed by David (exemplar breakpoint distance problem).
- First met David at APBC'07 in HK.
- First collaborated with David in 2010, on the scaffold filling problem.

1. David Sankoff's Contribution: My Personal Experience

- First collaborated with David in 2010, on the scaffold filling problem.

Munoz, Zheng, Q. Zhu, Albert, Rounsley, Sankoff. Scaffold filling, contig fusion and gene order comparison. BMC Bioinformatics 11, 2010.

1. David Sankoff's Contribution: My Personal Experience

- First collaborated with David in 2010, on the scaffold filling problem.

Munoz, Zheng, Q. Zhu, Albert, Rounsley, Sankoff. Scaffold filling, contig fusion and gene order comparison. BMC Bioinformatics 11, 2010.

Jiang, Zheng, Sankoff, B. Zhu. Scaffold filling under the breakpoint and related distances. IEEE/ACM TCBB 9, 2012.

Liu, Jiang, D. Zhu, B. Zhu. An improved approximation algorithm for scaffold filling to maximize the common adjacencies. IEEE/ACM TCBB, 2013 (published on-line on Aug 15, 2013).

2. The Exemplar Breakpoint Distance and Related Problems

- *In computational genomics, a lot of research has been performed on rearrangement for "ideal" genomes, i.e., permutations.*

2. The Exemplar Breakpoint Distance and Related Problems

- *In computational genomics, a lot of research has been performed on rearrangement for "ideal" genomes, i.e., permutations.*

For instance, the Sorting Signed Permutations by Reversals problem was shown to be in P (Hannenhalli and Pevzner, 1999); and Sorting by Transpositions problem was shown to be NP-hard recently (Bulteau et al., 2012).

2. The Exemplar Breakpoint Distance and Related Problems

- *In computational genomics, a lot of research has been performed on rearrangement for "ideal" genomes, i.e., permutations.*
- *However, due to the fast evolution/self-production, duplicated (paralogous) genes are common in some genomes. So it is important to select the ancestral ortholog of a gene family on an evolutionary basis.*

2. The Exemplar Breakpoint Distance and Related Problems

- *In computational genomics, a lot of research has been performed on rearrangement for "perfect" genomes, i.e., permutations.*
- *However, due to the fast evolution/self-production, duplicated (paralogous) genes are common in some genomes. So it is important to select the ancestral ortholog of a gene family on an evolutionary basis.*
- *In 1999, David Sankoff first formulated this as the exemplar breakpoint/genomic distance problem.*

2. The Exemplar Breakpoint Distance and Related Problems

- *Def. Given two permutations A and B over the same alphabet Σ, ab is a 2-substring in A but neither ab nor ba is a 2-substring in B, then ab is a breakpoint.*

2. The Exemplar Breakpoint Distance and Related Problems

- *Def. Given two permutations A and B over the same alphabet Σ, ab is a 2-substring in A but neither ab nor ba is a 2-substring in B, then ab is a breakpoint.*
- *Example. A = abcde, B = bcaed, then there are 2 breakpoints in A and B.*

2. The Exemplar Breakpoint Distance and Related Problems

- *Def. Given two permutations A and B over the same alphabet Σ, ab is a 2-substring in A but neither ab nor ba is a 2-substring in B, then ab is a breakpoint.*
- *If ab is a 2-substring in A and either ab or ba is a 2-substring in B, then ab is called an adjacency.*

Example. A = abcde, B = bcaed, then there are 2 adjacencies in A and B.

2. The Exemplar Breakpoint Distance and Related Problems

- *Problem: Given two genomes G' and H' with gene repetitions, compute two exemplar genomes G and H (i.e., exactly one gene in each family is kept) such that the number of breakpoints (resp. adjacencies) between G and H is minimized (resp. maximized).*

2. The Exemplar Breakpoint Distance and Related Problems

- *Problem: Given two genomes G' and H' with gene repetitions, compute two exemplar genomes G and H (i.e., exactly one gene in each family is kept) such that the number of breakpoints (resp. adjacencies) between G and H is minimized (resp. maximized).*
- *Example: G'=badcbda, H'=abcdab*

  optimal:

2. The Exemplar Breakpoint Distance and Related Problems

- *Problem: Given two genomes G' and H' with gene repetitions, compute two exemplar genomes G and H (i.e., exactly one gene in each family is kept) such that the number of breakpoints (resp. adjacencies) between G and H is minimized (resp. maximized).*
- *Example: G'=badcbda, H'=abcdab*

  optimal: G = bcda, H = bcda

  # breakpoints = 0, # adjacencies = 3

2. The Exemplar Breakpoint Distance and Related Problems

- *David Bryant proved that the Exemplar Breakpoint Distance problem is NP-complete in 2000.*
- *In 2005, I ran a workshop with Zhixiang Chen and Bin Fu and we proved that the Exemplar Breakpoint Distance problem does not admit any polynomial-time approximation, unless P=NP, even when each gene appears at most three times (Chen,Fu,Zhu, AAIM'06). (Improved to 2-times, a few years later by Angibaud et al. 2009; Jiang, 2010).*

2. The Exemplar Breakpoint Distance and Related Problems

- *3SAT < ZERO-EBD*

Example. $\Phi = F_1 \wedge F_2 \wedge F_3 \wedge F_4$, *where* $F_1 = x_1 \vee \neg x_2 \vee x_3$,

$F_2 = \neg x_1 \vee x_2 \vee \neg x_4$, $F_3 = \neg x_2 \vee \neg x_3 \vee x_4$, $F_4 = x_1 \vee \neg x_3 \vee \neg x_4$.

For each x_i, *define* S_i *(resp.* S_i'*) as the list of clauses containing* x_i *(resp.* $\neg x_i$*) followed by the clauses containing* $\neg x_i$ *(resp.* x_i*).*

Example. $S_1 = F_1 F_4 F_2$, $S_1' = F_2 F_1 F_4$.

2. The Exemplar Breakpoint Distance and Related Problems

- *3SAT < ZERO-EBD*

Example. $\Phi=F_1 \wedge F_2 \wedge F_3 \wedge F_4$, *where* $F_1=x_1 \vee \neg x_2 \vee x_3$,

$F_2=\neg x_1 \vee x_2 \vee \neg x_4$, $F_3=\neg x_2 \vee \neg x_3 \vee x_4$, $F_4=x_1 \vee \neg x_3 \vee \neg x_4$.

Construct two genomes

$G'=S_1 g_1 S_2 g_2 S_3 g_3 S_4$, $H'=S_1' g_1 S_2' g_2 S_3' g_3 S_4'$.

If x_i*=True then keep the clauses in* S_i *and* S_i' *containing* x_i *and vice versa, then delete the remaining duplicated clauses arbitrarily.*

2. The Exemplar Breakpoint Distance and Related Problems

- *3SAT < ZERO-EBD*

Example. $\Phi=F_1 \wedge F_2 \wedge F_3 \wedge F_4$ *, where* $F_1=x_1 \vee \neg x_2 \vee x_3$ *,*

$F_2=\neg x_1 \vee x_2 \vee \neg x_4$ *,* $F_3=\neg x_2 \vee \neg x_3 \vee x_4$ *,* $F_4=x_1 \vee \neg x_3 \vee \neg x_4$ *.*

Construct two genomes

$G'=S_1 g_1 S_2 g_2 S_3 g_3 S_4$, $H'=S_1' g_1 S_2' g_2 S_3' g_3 S_4'$.

With this example, we can obtain $G=H=F_4 g_1 F_3 g_2 F_1 g_3 F_2$ *(*$d(G,H)=0$*) with* $x_1=x_3=$*True and* $x_2=x_4=$*False.*

2. The Exemplar Breakpoint Distance and Related Problems

- *3SAT < ZERO-EBD*
- *The construction is simple and can easily produce sequences for NP-hardness proofs in various applications, e.g., computational geometry, protein structure simplification, and multi-channel program downloading.*
- *The 2-repetition construction by Angibaud et al. and Jiang is still too complex to have extra applications.*

2. The Exemplar Breakpoint Distance and Related Problems

- *3SAT < ZERO-EBD*
- *Implications:*

(1)EBD has no polynomial time approximation unless P=NP.

(2)EBD has no FPT algorithm unless P=NP.

These results hold even when a gene appears at most twice.

2. The Exemplar Breakpoint Distance and Related Problems

- *Implications:*

(1)EBD has no polynomial time approximation unless P=NP.

(2)EBD has no FPT algorithm unless P=NP.

Open Problem #1: What if one of the two input genomes is exemplar, i.e., what is the approximability of the One-sided EBD?

2. The Exemplar Breakpoint Distance and Related Problems

Open Problem #1: What if one of the two input genomes is exemplar, i.e., what is the approximability of the One-sided EBD?

Status: NP-hard and APX-hard, the only known approximation bound is Θ(n).

2. The Exemplar Breakpoint Distance and Related Problems

- *For the dual problem of EBD:*

  Independent Set < Exemplar Adjacency (Chen et al., CPM'07)

The Exemplar Adjacency problem does not admit any polynomial-time factor $n^{0.5-\varepsilon}$ approximation unless NP=ZPP (and, no FPT algorithm unless FPT=W[1]). This holds even when one genome is exemplar and each gene in the other appears at most twice. Moreover, there are matching approximations.

2. The Exemplar Breakpoint Distance and Related Problems

Problem	Inapproximability	FPT Tractability
Exemplar Breakpoint Distance	No poly-time approximation, unless P=NP	No FPT algorithm, unless P=NP
Exemplar Adjacency	Can't have a factor better than $n^{0.5-\varepsilon}$, unless NP=ZPP	No FPT algorithm, unless FPT=W[1]

3. Maximal Strip Recovery and Its Complement

- Given two comparative maps, with gene markers, we want to identify noise and redundant markers. *Note that in comparative maps only the relative positions of the markers along chromosome are indicated (Bertrand, Blanchette, El-Mabrouk, 2009, …).*
- In 2007, David Sankoff first formalized this as an algorithmic problem *(Zheng, Q.Zhu, Sankoff, TCBB, 2007).*

3. Maximal Strip Recovery and Its Complement

Example.

G_1=<1,2,3,4,5,6,7,8,9,10,11,12>

G_2=<-8,-5,-7,-6,4,1,3,2,-12,-11,-10,9>

Given two comparative maps, with gene markers, we want to identify noise and redundant markers.

3. Maximal Strip Recovery and Its Complement

Example.

G_1=<1,2,3,4,5,6,7,8,9,10,11,12>

G_2=<-8,-5,-7,-6,4,1,3,2,-12,-11,-10,9>

G'_1=<1,3,6,7,8,10,11,12>

G'_2=<-8,-7,-6,1,3,-12,-11,-10>

This can be done by first finding syntenic blocks (strips) with maximum total length.

3. Maximal Strip Recovery and Its Complement

Example.

G_1=<1,2,3,4,5,6,7,8,9,10,11,12>

G_2=<-8,-5,-7,-6,4,1,3,2,-12,-11,-10,9>

G'_1=<1,3,6,7,8,10,11,12>, 3 syntenic blocks

G'_2=<-8,-7,-6,1,3,-12,-11,-10>

G_1=<1,2,3,4,5,6,7,8,9,10,11,12>, redundant

G_2=<-8,-5,-7,-6,4,1,3,2,-12,-11,-10,9>

3. Maximal Strip Recovery and Its Complement

- A strip (syntenic block) is a string of distinct markers that appear in two maps, either directly or in reversed and negated form.

  Example. 6,7,8 in G'_1; -8,-7,-6 in G'_2.

- MSR (Maximal Strip Recovery): Given two maps G and H, find two subsequences G' and H' of G and H, such that the total length of disjoint strips in G' and H' is maximized.

- CMSR --- complement of MSR.

3. Maximal Strip Recovery and Its Complement

- After some struggle, both MSR and CMSR were shown to be NP-complete (Wang, Zhu, JCB 2010).
- The approximation and FPT algorithm research attracted 2 additional groups in Canada and France.

3. Maximal Strip Recovery and Its Complement

Problem	Approximability	FPT Tractability
MSR	Factor 4	?
CMSR	Factor 3, then 2.33	$O^*(2.36^k)$ Best kernel: 78k

4. Scaffold Filling Problems

- *Motivation: Most of the genomes sequenced are not really sequenced, they are typically presented as scaffolds.*

4. Scaffold Filling Problems

- *Motivation: Most of the genomes sequenced are not really sequenced, they are typically presented as scaffolds.*
- *For a singleton genome, possibly with gene repetitions, a scaffold is simply an incomplete sequence (with some genes missing).*

4. Scaffold Filling Problems

- *Motivation: Most of the genomes sequenced are not really sequenced, they are typically presented as scaffolds.*
- *For a singleton genome, possibly with gene repetitions, a scaffold is simply an incomplete sequence (with some genes missing).*

Example: G=#abcdefg# is a complete reference genome, H=#gbcdf# is a scaffold with gene a,e missing.

4. Scaffold Filling Problems

- *For a singleton genome, possibly with gene repetitions, a scaffold is simply an incomplete sequence (with some genes missing).*

Example: G=#abcdefg# is a complete reference genome, H=#gbcdf# is a scaffold with gene a,e missing.

Problem: Insert the missing genes into H to obtain H' s.t. d(G,H') is small or some similarity between G and H' is large.

4. Scaffold Filling Problems

Status:

- *When there is no gene repetition, the one-sided problem under the DCJ distance (for multichromosome genomes) is in P (Munoz et al. 2010).*
- *When there is no gene repetition, even the two-sided problems are in P (Jiang et al. 2012).*
- *When there are gene repetitions, the one-sided problem is NP-hard, using string adjacencies as a similarity measure (Jiang et al. 2012).*

4. Scaffold Filling Problems

Definition:

- *Given a scaffold/sequence A, P_A represents all A's 2-substrings.*
- *Given scaffolds/sequences A and B, if a 2-substring of A, $a_i a_{i+1}$, is equal to $b_j b_{j+1}$ or $b_{j+1} b_j$ then we say that they are matched to each other. In a maximum matching of the pairs in P_A and P_B, a matched pair is called an adjacency, and any unmatched pair is called a breakpoint.*

4. Scaffold Filling Problems

Definition:

- *Given scaffolds/sequences A and B, if a 2-substring of A, $a_i a_{i+1}$, is equal to $b_j b_{j+1}$ or $b_{j+1} b_j$ in B then we say that they are matched to each other. In a maximum matching of the pairs in P_A and P_B, a matched pair is called an adjacency, and any unmatched pair is called a breakpoint.*

Example: A=#13245#,

B=#12356#, adjacencies: #1, 23

breakpoints in A: 13, 24, 45, 5#;

breakpoints in B: 12, 35, 56, 6#

4. Scaffold Filling Problems

Note that (string) adjacencies and breakpoints are completely different from those for permutations.

For instance, even if there is no breakpoint we could have A ≠ B or its reversal.

Example: A=#bcdabc#,

B=#bcbadc#, there is no breakpoint.

But A ≠ B or its reversal.

4. Scaffold Filling Problems

Note that (string) adjacencies and breakpoints are completely different from those for permutations.

For instance, even if there is no breakpoint we could have A ≠ B or its reversal.

Example: A=#bcdabc#,

B=#bcbadc#, there is no breakpoint.

But A ≠ B or its reversal.

Nevertheless, biologically and intuitively, more adjacencies would be good.

4. Scaffold Filling Problems

Approximation Results:

(1)For the one-sided scaffold filling to maximize the number of common adjacencies problem, there is a factor 1.33 approximation (Jiang, Zheng, Sankoff, B. Zhu, IEEE/ACM TCBB 9, 2012.)

4. Scaffold Filling Problems

Approximation Results:

(1)For the one-sided scaffold filling to maximize the number of common adjacencies problem, there is a factor 1.33 approximation (Jiang, Zheng, Sankoff, B. Zhu, IEEE/ACM TCBB 9, 2012.)

Method: greedy search.

4. Scaffold Filling Problems

Approximation Results:

(1)For the one-sided scaffold filling to maximize the number of common adjacencies problem, there is a factor 1.33 approximation (Jiang, Zheng, Sankoff, B. Zhu, IEEE/ACM TCBB 9, 2012).

(2)There is a factor 1.25 approximation (Liu, Jiang, D. Zhu, B. Zhu, IEEE/ACM TCBB, 2013).

Method: maximum matching, local search and greedy search.

(*) My Collaborators

- *Zhixiang Chen, Richard Fowler, Bin Fu (U. of Texas-Pan American).*
- *Haitao Jiang, Nan Liu, Daming Zhu (Shandong University, China).*
- *Zhong Li, Guohui Lin, Weitian Tong (University of Alberta).*
- *David Sankoff, Chunfang Zheng (University of Ottawa).*
- *Minghui Jiang, Lusheng Wang, Boting Yang*

(**) Tenure Track Openings at Montana State University

- *At least 2 openings in CS, ad will be out soon.*
- *Bioinformatics is one of the targeted areas.*
- *Female candidates are especially welcome.*
- *City Bozeman: 40K population, near Yellowstone, constantly ranked "best town to live" in US, especially suitable for people loving outdoor activities …*