

Technical Report:
Assessing Method for *de novo* Repeat Identification

Zhirong Bao and Sean R. Eddy¹

Howard Hughes Medical Institute and

Department of Genetics

Washington University School of Medicine

St. Louis, MO 63110 USA

¹Tel: +1 314 362 7666; Fax: +1 314 362 7855; Email: eddy@genetics.wustl.edu.

We assess the result of a *de novo* repeat family identification by comparing it to a trusted result. Comparing two results of repeat family identification is similar to comparing two results of gene prediction, with elements corresponding to exons and families corresponding to genes. However, the former can be more complicated, as one nucleotide position may be assigned simultaneously to multiple elements and families. Here, we present our method of comparing two results of repeat family identification, which was used to train the RECON package.

Notations and Definitions

Let $\{F_\alpha = \{E_i^\alpha\}\}$ denote the true MCS families and the copies/elements of each family in the input sequences, and $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$ denote the families and their copies identified by a given method. Also, let $\{E_i\}$ and $\{\mathcal{E}_j\}$ denote the set of true and identified elements (regardless of which family they belong to).

A true element E_i and an identified element \mathcal{E}_j are considered to correspond to each other if the overlap between the two is longer than 50% of either of the elements. A true family F_α and an identified family \mathcal{F}_β are considered to correspond to each other if any E_i^α corresponds to any \mathcal{E}_j^β . Let $m(E_i, \mathcal{E}_j) = 1$ if E_i and \mathcal{E}_j correspond and $m(E_i, \mathcal{E}_j) = 0$ if E_i and \mathcal{E}_j do not correspond.

The Strategy

Our goal is to quantitate the difference between $\{F_\alpha = \{E_i^\alpha\}\}$ and $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$. The comparison would be straightforward if a nucleotide position can be assigned to at most one \mathcal{E}_j and there is one-to-one/zero correspondence between $\{E_i\}$ and $\{\mathcal{E}_j\}$ and between $\{F_\alpha = \{E_i^\alpha\}\}$ and $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$. Unfortunately, it is usually not the case in practice.

We capture the following three types of differences/errors that could occur in $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$ when compared to $\{F_\alpha = \{E_i^\alpha\}\}$:

1. lack of sensitivity, including: (a) failing to identify an element, i.e., for an $E \in \{E_i\}$, there is no $\mathcal{E} \in \{\mathcal{E}_j\}$ corresponding to it; (b) failing to identify part of an element, i.e., certain nucleotide positions in an $E \in \{E_i\}$ are not in its corresponding $\mathcal{E} \in \{\mathcal{E}_j\}$; and (c) breaking a true family into several identified families, i.e., \mathcal{E}_j 's that correspond to $\{E_i^\alpha\}$ are assigned to different \mathcal{F}_β 's;
2. redundancy, i.e., simultaneously assigning a position in $\{F_\alpha = \{E_i^\alpha\}\}$ to multiple \mathcal{F}_β 's;
3. lack of specificity, including: (a) identifying an $\mathcal{E} \in \{\mathcal{E}_j\}$ that does not correspond to any $E \in \{E_i\}$; (b) some positions in an $\mathcal{E} \in \{\mathcal{E}_j\}$ not in its corresponding $E \in \{E_i\}$; and (c) lumping families, i.e., \mathcal{E}_j 's in an \mathcal{F}_β corresponding to true elements that are in different F_α 's.

This is of course not a complete list of possible differences/errors, but rather those that could potentially affect the quality of the sequence models built on top of each $\{\mathcal{F}_\beta = \{\mathcal{E}_j^\beta\}\}$. To quantify, we count the amount of nucleotide positions involved in each of the three types of errors.

The Formula

For a set of nucleotide positions E , we define $|E|$ as the number of nucleotide positions in E . For a set of elements $F = \{E_i\}$, we define $|F| = |\bigcup_{E_i \in F} E_i|$, which is the number of nucleotide positions in F with redundant positions being counted

only once. Furthermore, we define

$$u(F_\alpha, \mathcal{F}_\beta) = u(\mathcal{F}_\beta, F_\alpha) = \bigcup_{m(E_i^\alpha, \mathcal{E}_j^\beta)=1} E_i^\alpha \cap \mathcal{E}_j^\beta$$

which is the set of nucleotide positions in F_α which are properly identified in \mathcal{F}_β .

Also, let $\mathcal{B}(F_\alpha)$ denote F_α 's best match in $\{\mathcal{F}_\beta\}$, so that

$$\mathcal{B}(F_\alpha) = \arg \max_{\{\mathcal{F}_\beta\}} \left(\left| u(F_\alpha, \mathcal{F}_\beta) \right| \right),$$

or $\mathcal{B}(F_\alpha) = \phi$ if F_α does not match any \mathcal{F}_β . Similarly, let $B(\mathcal{F}_\beta)$ denote \mathcal{F}_β 's best match in $\{F_\alpha\}$, so that

$$B(\mathcal{F}_\beta) = \arg \max_{\{F_\alpha\}} \left(\left| u(\mathcal{F}_\beta, F_\alpha) \right| \right),$$

or $B(\mathcal{F}_\beta) = \phi$ if \mathcal{F}_β does not match any F_α .

For a given true family F_α , the error for the lack of sensitivity (denoted by Err_{1, F_α}) is calculated as

$$\begin{aligned} Err_{1, F_\alpha} &= \left| F_\alpha \right| - \left| \bigcup_{\{\mathcal{F}_\beta\}} u(F_\alpha, \mathcal{F}_\beta) \right| \\ &\quad + \left| \bigcup_{\{\mathcal{F}_\beta\}} u(F_\alpha, \mathcal{F}_\beta) \right| - \left| u(F_\alpha, \mathcal{B}(F_\alpha)) \right|. \end{aligned}$$

The first two terms measure error 1(a) and 1(b). The last two terms measure error 1(c) by counting the number of identified nucleotide positions in F_α which are not found in its best match $\mathcal{B}(F_\alpha)$. After canceling the second and the third terms,

$$Err_{1, F_\alpha} = \left| F_\alpha \right| - \left| u(F_\alpha, \mathcal{B}(F_\alpha)) \right|$$

which is equivalent to the number of nucleotide positions in F_α which are not properly identified in $\mathcal{B}(F_\alpha)$.

The error of redundancy concerning F_α (Err_{2,F_α}) is calculated as

$$Err_{2,F_\alpha} = \sum_{\{\mathcal{F}_\beta\}} \left| u(F_\alpha, \mathcal{F}_\beta) \right| - \left| \bigcup_{\{\mathcal{F}_\beta\}} u(F_\alpha, \mathcal{F}_\beta) \right|.$$

For a given identified family \mathcal{F}_β , the error for the lack of specificity ($Err_{3,\mathcal{F}_\beta}$) is calculated as

$$Err_{3,\mathcal{F}_\beta} = \left| \mathcal{F}_\beta \right| - \left| u(\mathcal{F}_\beta, B(\mathcal{F}_\beta)) \right|.$$

The total error (Err) is defined as

$$Err = \sum_{\{F_\alpha\}} Err_{1,F_\alpha} + \sum_{\{F_\alpha\}} Err_{2,F_\alpha} + \sum_{\{\mathcal{F}_\beta\}} Err_{3,\mathcal{F}_\beta} \quad (1)$$

For Optimization

Typically, we know the representative sequence of certain families in the subject genome. We can carefully locate copies of these families in the input sequences, and optimize the whole analysis by reducing the errors concerning these families/copies. Accordingly, in equation 1, the first two terms will sum over these known families, and the last term will sum over identified families which correspond the known families, using the corresponding known family as the best match for the identified families concerned.