Algebraic studies of giant chromosomes in genus Chironomus

Szabolcs Surányi

The genus Chironomus (midge) is widely variable genetically due to frequent mutations. The Chironomus species can be identificated by the banding patterns of the giant chromosomes in the salivary gland cells of the larvae, which form groups in 7 arms. The banding patterns consist of atomic and unique bands. In practice, band sequences change in a very special way during mutation: analysis of band sequences of hundreds of species shows that only so called inversion produces new species. Specifically, when inversion occurs during the copy process, some continuous part of the sequence is cut out, turned around and sticked back in its original position but in opposite direction.

Given the band sequences of about $100 - 200$ species, the main problem is to produce the most likely philogenetic tree for the species. The most likely philogenetic tree is a directed graph, such that every vertex represents a species, and every edge points from a species to its most likely ancestor species. Obviously, to solve this problem it is inevitable to have a method by which we can find quickly all possible inversion series, which could produce a given species from another.

To deal with the notions and problems mentioned above, we have created a mathematical model based on the theory of symmetric groups. Arms of n bands are represented as permutations of S_n, and we define the set $P(S_n) = \{1, 2, \ldots, n, n+1\}$ as the points of S_n, which can be the possible locations, where the sequence is cut, as shown below:

$$\psi = \left(\begin{array}{ccccccccccc} | & 1 & | & 2 & | & 3 & | & \ldots & | & n & | \\ \scriptstyle 1 & & \scriptstyle 2 & & \scriptstyle 3 & & \scriptstyle 4 & & \scriptstyle \mathrm{n} & & \scriptstyle \mathrm{n+1} \\ | & i_1 & | & i_2 & | & i_3 & | & \ldots & | & i_n & | \end{array}\right).$$

Thus the $[p, q]$ inversion between the $p, q \in P(S_n)$, $p \leq q$ points is defined as follows: if $p = q$ then $[p, q]$ is the identical element, otherwise it is the following permutation:

$$[p, q] = \begin{pmatrix} 1 & 2 & \ldots & p-1 & p & p+1 & \ldots & q-2 & q-1 & q & \ldots & n-1 & n \\ 1 & 2 & \ldots & p-1 & q-1 & q-2 & \ldots & p+1 & p & q & \ldots & n-1 & n \end{pmatrix}.$$

Using these notations we say that a $[p_1, q_1], [p_2, q_2], \ldots, [p_k, q_k]$ $(p_i, q_i \in P(S_n))$ inversion series of length k is a derivation from $\psi \in S_n$ to $\phi \in S_n$, if $\phi = \psi \cdot \prod_{i=1}^{k} [p_i, q_i]$.

It is assumed that in nature superfluous inversions does not occur. According to this assumption, our problem can be described as a brief question: For given $\psi, \phi \in S_n$ permutations how can we find a derivation from ψ to ϕ of minimal length.

We have proved that for arbitrary $\psi, \phi \in S_n$ a derivation from ψ to ϕ of length n can be found, which on the other hand proves that the set of inversions $I \subseteq S_n$ generate S_n. In practice this bound can be improved such a way that the bound is not a function of n, but m, where m is the number of so called breakpoints in ψ according to ϕ.

The set of breakpoints in ψ according to ϕ can be defined based on neighbourhood of elements: The left neighbourhood of point $p \in P(S_n)$ in $\psi \in S_n$ denoted by $N_L(\psi, p)$ is 0 if $p = 1$, $\psi(p-1)$ otherwise. Analogously, the right neighbourhood of point $p \in P(S_n)$ in $\psi \in S_n$ denoted by $N_R(\psi, p)$ is $n+1$ if $p = n+1$, $\psi(p)$ otherwise. Thus, the neighbourhood of point $p \in P(S_n)$ in $\psi \in S_n$ is the set $\{N_L(\psi, p), N_R(\psi, p)\}$. Using these definitions the set of breakpoints in $\psi \in S_n$ according to $\phi \in S_n$ can be defined as the set $B(\psi, \phi) = \{p \in P(S_n) \mid \neg\exists q \in P(S_n) \text{ such that } N(\phi, p) = N(\psi, p)\}$.

We have given an algorithm, which produces for arbitrary $\psi, \phi \in S_n$ a derivation from ψ to ϕ of length m, where $m = |B(\psi, \phi)|$. Considering that $m \leq n+1$, this bound gives us a much better bound in almost every case, furthermore we have proved that it is a better bound in every case.

Besides these results we have proved the existence of a lower bound, which is $\lceil\ |B(\psi,\phi)|\ /\ 2\ \rceil$ for a given $\psi,\phi\ \in\ S_n$, thus the number of inversions required to derive a species from another is squeezed between quite strict bounds.

After sending the abstract it turned out that these results are rather old, refer to John D. Kececioglu and David Sankoff, Exact and Approximation Algorithms for the Inversion Distance Between Two Chromosomes, Lecture Notes In Computer Science, 87-105, 1993.