DESCRIPTION:

The appendix to the paper "Longest Order Conserved Exemplar Subsequences".

Lemmas:

Lemma 1. *If for $i \geq x_k$ (resp. $j \geq y_k$) where $1 \leq k \leq q$, a member in $C(i,\, j)$ fails to be indexed by $X[1, k]$ as well as $Y[1, k]$, then any extension of the member cannot be indexed by X as well as Y.*

Proof. Let $C_1 \in C(i, j)$, $C \in C(m, n)$. If C is an extension of C_1, then there is a repetition-free common subsequence C_2 of $A[i + 1,\, m]$ and $B[j + 1,\, n]$, such that $C = C_1 \parallel C_2$.

It follows from $occ(A,\, A[x_p]) = 1$ for $1 \leq p \leq k$ that C_2 cannot be indexed by $X[p]$ for $1 \leq p \leq k$. If C_1 fails to be indexed by $X[1,\, k]$ as well as $Y[1,\, k]$, so does C.

Lemma 2. *For k with $1 \leq k \leq q$, a repetition-free common subsequence $C \in C(x_k,\, y_k)$ is indexed by $X[1,\, k]$ as well as $Y[1,\, k]$, if and only if there exists C' $\in C(x_k - 1,\, y_k - 1])$ indexed by $X[1,\, k - 1]$ as well as $Y[1,\, k - 1]$ such that C $= C' \parallel A[x_k]$.*

Proof. **If**: Let $C' \in C(x_k - 1,\, y_k - 1)$. It follows from $occ(A,\, A[x_k]) = occ(B,$ $B[y_k]) = 1$ that no gene identical to $A[x_k]$ or $B[y_k]$ can occur in C'. It follows from $A[x_k] = B[y_k]$ that $C' \parallel A[x_k] \in C(x_k,\, y_k)$ and if C' is indexed by $X[1,$ $k - 1]$ as well as $Y[1,\, k - 1]$, then $C' \parallel A[x_k]$ is indexed by $X[1,\, k]$ as well as $Y[1,\, k]$.

Only if: Let $C \in C(x_k,\, y_k)$ be indexed by $X[1,\, k]$ as well as $Y[1,\, k]$. Since $A[x_k] = B[y_k]$ and $occ(A,\, A[x_k]) = occ(B,\, B[y_k]) = 1$, C can be expressed as C' $\parallel A[x_k]$ where $C' \in C(x_k - 1,\, y_k - 1)$. Since C is indexed by $X[1,\, k]$ as well as $Y[1,\, k]$, C' must be indexed by $X[1,\, k - 1]$ as well as $Y[1,\, k - 1]$.

Lemma 3. *Let for i and j with $x_k \leq i < x_{k+1}$ and $y_k \leq j < y_{k+1}$ where 0 $\leq k \leq q$, $C_1 \in CP(i,j)$ and $C_2 \in CP(i,j)$. If $f(i,\, j,\, C_1) = f(i,\, j,\, C_2)$ and $|C_1| \geq |C_2|$, then in $CP(m,n)$, a longest extension of C_1 is no shorter than any extension of C_2.*

Proof. Let C be a longest extension of C_2 in $CP(m,\, n)$. Then there exists a repetition-free common subsequence of $A[i + 1,\, m]$ and $B[j + 1,\, n]$ indexed by $X[k + 1,\, q]$ as well as $Y[k + 1,\, q]$, say C_3 such that $C = C_2 \parallel C_3$. Then it follows from $f(i,\, j,\, C_1) = f(i,\, j,\, C_2)$ and C_1 is indexed by $X[1,\, k]$ as well as $Y[1,\, k]$ that $C_1 \parallel C_3$ is a repetition-free common subsequence of A and B indexed by X as well as Y that is an extension of C_1 in $C(m, n)$. Since $|C_1| \geq |C_2|$, $|C_1 \parallel$ $C_3| \geq |C|$. The lemma follows from that the longest extension of C_1 in $C(m, n)$ has no less genes than $|C_1 \parallel C_3|$.

Lemma 4. *If $A[i] = B[j]$ for i and j with $i + j > 0$, $x_k \leq i < x_{k+1}$ and $y_k \leq j < y_{k+1}$ $(0 \leq k \leq q)$, then $CFP = \{C' \parallel A[i] : C' \in CFP(i-1, j-1, \overline{A[i]})\}$ is a minimum representative subset of $CP(i, j, A[i])$.*

Proof. Then we argue for CFP to be representative in the following two aspects.

(1)Let $C'_1 \parallel A[i] \in CP(i, j, A[i])$ where $C'_1 \in CP(i-1, j-1, \overline{A[i]})$. Then since $CFP(i-1, j-1, \overline{A[i]})$ is representative, there exists a $C' \in CFP(i-1, j-1, \overline{A[i]})$ with $f(i-1, j-1, C') = f(i-1, j-1, C'_1)$. Then $f(i, j, C' \parallel A[i]) = f(i, j, C'_1 \parallel A[i])$.

(2)Let $C' \parallel A[i] \in CFP$, $C'_1 \parallel A[i] \in CP(i, j, A[i])$. If $f(i, j, C'_1 \parallel A[i]) = f(i, j, C' \parallel A[i])$, then since $C' \in CFP(i-1, j-1, \overline{A[i]})$, $C'_1 \in CP(i-1, j-1, \overline{A[i]})$, $f(i-1, j-1, C') = f(i-1, j-1, C'_1)$. It follows from $|C'_1| \leq |C'|$ that $|C'_1 \parallel A[i]| \leq |C' \parallel A[i]|$.

The reason why CFP is minimum over all representative subsets of $CP(i, j, A[i])$ lies in that $CFP(i-1, j-1, \overline{A[i]})$ is minimum over all those representative subsets of $CP(i-1, j-1, \overline{A[i]})$.

Lemma 5. *For i and j with $x_k \leq i < x_{k+1}$ and $y_k \leq j < y_{k+1}$ $(0 \leq k \leq q)$, $|FP(i,j)| \leq 2^{min\{s(A),s(B)\}}$.*

Proof. Without loss of generality, let $s(A) = min\{s(A), s(B)\}$. Then there are at most $s(A)$ gene families in both $A[1, i]$ and $A[i+1, m]$ for any i with $0 \leq i \leq m$. A confused gene family of an arbitrary member in $CP(i,j)$ must occur in both $A[i+1{,}m]$ and $B[j+1, n]$. Since at most $s(A)$ gene families can occur in both $A[1, i]$ and $A[i+1, m]$, an arbitrary member in $CP(i, j)$ can admit a confused gene family set of at most $s(A)$ gene families. The lemma follows from that those confused gene family sets in $FP(i, j)$ are mutually different.

Tables:

Table 1. Lengths of human/gorilla chromosome summaries

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9
human	5475	4200	3188	2657	2988	3064	3014	2485	2333
gorilla	2947	2022	1716	1223	2094	1542	1440	1107	1165
	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18
human	2336	3364	3055	1402	2287	2219	2558	3059	1244
gorilla	1127	1728	1506	611	1102	991	1144	820	474
	chr19	chr20	chr21	chr22	chrX				
human	2991	1458	875	1388	2425				
gorilla	1668	787	307	645	1315				

A "huamn" or "gorilla" statistic represents the gene number of a human or gorilla chromosome summary.

.

Table 2. RFLCS length ratios of human/gorilla chromosome summary for 23 pairs

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX
rflcs/lh	0.303	0.156	0.299	0.225	0.156	0.279	0.199	0.157	0.204	0.185	0.316	0.177	0.200	0.187	0.212	0.260	0.076	0.144	0.372	0.318	0.183	0.258	0.269
rflcs/lg	0.563	0.325	0.556	0.489	0.223	0.554	0.416	0.353	0.409	0.382	0.615	0.360	0.460	0.388	0.474	0.580	0.282	0.378	0.668	0.588	0.521	0.555	0.496

A ”rflcs/lh” (resp. ”rflcs/lg”) statistic represents the ratio of the gene number of a RFLCS to the gene number of a human (resp. gorilla) chromosome summary.

Table 3. Lengths of 46 chromosomes

	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX
Human	7715	11718	5907	6480	2661	5060	5888	5425	6970	5332	4141	5763	2595	3785	2981	3449	473	2791	1784	1672	1045	1360	5352
Gorilla	7296	8117	6184	5944	2820	5149	5199	4273	4864	4313	4145	4249	2708	2844	2863	3082	500	2293	1771	1702	1053	1331	5403

A ”Human” or ”Gorilla” statistic represents the length of a human or gorilla pseudo gene summary.