Hierarchical Clustering with Membrane Computing

M. Cardona, M. A. Colomer, M. J. Pérez-Jiménez,

A. Zaragoza

Universidad de Lleida, Universidad de Sevilla (Spain)

27th June 2007

The main Goal is:

- Classify a group of individuals with dichotomizing variables (measuring the presence or absence of certain qualitative characteristics).

- Design a P system for each Boolean matrix associated with a group of individuals.

Hierarchical Clustering

- The clustering is used to characterize and to order a vast amount of information on the variability of population of individuals.
- It is a technique based in statistical methods.
- It is obtained a grouping formed by similar individuals (cluster).
- The result is a hierarchy where the individuals are grouped in levels. Each level is a partition of the set of the individuals.

The individuals to classify are the elements of the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$

Each individual are k-tuple formed by *k* characteristics or variables.

The values of the individuals can be described in the matrix form:

$$P_{Nk} = \begin{pmatrix} \omega_{11} & \omega_{12} & \ldots & \omega_{1k} \\ \omega_{21} & \omega_{22} & \ldots & \omega_{2k} \\ & & \ldots & \\ \omega_{N1} & \omega_{N2} & \ldots & \omega_{Nk} \end{pmatrix} \qquad \boldsymbol{\omega_{ij} \in \{0,1\} \quad 1 \le i, j \le k}$$

To make the hierarchy we need:

- A function measuring the similarity between individuals. We use ***the similarity*** defined as follows:

$$s(\omega_i, \omega_j) = \sum_{r=1}^{k} \left(1 - \left|\omega_{ir} - \omega_{jr}\right|\right)$$

- A function measuring the similarity between clusters, called aggregation index. We use ***the aggregation index*** based on the minimum defined by:

$$\delta(h, h') = min\left\{s(\omega_i, \omega_j) \middle| \omega_i \in h, \omega_j \in h'\right\} \quad h, h' \in P(\Omega)$$

- A function measuring the homogeneity degree between the individuals belonging to the same cluster. This application is called ***hierarchical index***.

$$f(h) = \delta(h_1, h_2) \quad if \quad h = h_1 \cup h_2,\ h, h_1, h_2 \in P(\Omega)$$

An algorithm for the construction of a hierarchy:

1. The first level of the hierarchy :

$$P_0 = \{ S_1 = \{\omega_1\}, S_2 = \{\omega_2\}, \ldots, S_N = \{\omega_N\}\}$$

the *aggregation index* is

$$\delta(\{\omega_i\},\{\omega_j\}) = s(\omega_i, \omega_j)$$

and the *hierarchical index* is

$$f(\{\omega_i\}) = k \quad 1 \le i \le N$$

2. Find the two closest clusters S_i, S_j $\quad 1 \leq i < j \leq N$

and we consider a new cluster $S_i = S_i \cup S_j$

- Remove S_j from P_0, and the new partition is:

$$P_1 = \{S_1, \ldots, S_i = \{S_i \cup S_j\}, \ldots, S_{j-1}, S_{j+1}, \ldots S_N\}$$

3. Compute *the aggregation index* between all pairs of the clusters in the new partition.

4. Go to step 2 until there is only one set remaining.

Remark: If at step 2 there are more than one possibility, then one of them is chosen so the hierarchy obtained is not unique.

Example of construction of the indexed hierarchy

1- The individuals to classify are

$$\omega_1 = (1,0,0,0)\ \omega_2 = (0,1,1,0)\ \omega_3 = (1,1,1,0)\ \omega_4 = (1,1,1,1)$$

- The matrix of similarities is:

$$\begin{pmatrix} 4 & 1 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Example of construction of the indexed hierarchy

1- We want to classify the individuals

$\omega_1 = (1,0,0,0)$ $\omega_2 = (0,1,1,0)$ $\omega_3 = (1,1,1,0)$ $\omega_4 = (1,1,1,1)$

- The matrix of similarities is:

$$\begin{pmatrix} 4 & 1 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

$$s(\omega_1, \omega_3) = (1-|1-1|) + (1-|0-1|) + (1-|0-1|) + (1-|0-0|)$$

Example of construction of the indexed hierarchy

1- We want to classify the individuals

$$\omega_1 = (1,0,0,0)\ \omega_2 = (0,1,1,0)\ \omega_3 = (1,1,1,0)\ \omega_4 = (1,1,1,1)$$

- The matrix of similarities is:

$$\begin{pmatrix} 4 & 1 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

- The initial partition is $P_0 = \{\{\omega_1\},\{\omega_2\},\{\omega_3\},\{\omega_4\}\}$

- The hierarchical index are

$$f(\{\omega_1\}) = f(\{\omega_2\}) = f(\{\omega_3\}) = f(\{\omega_4\}) = 4$$

- The aggregation index in this step is the same as the similarity

2- *We obtain the partition P_1 joining the two most similar clusters of the previous partition. Maximizing δ*

$$\begin{pmatrix} 4 & 1 & 2 & 1 \\ 1 & 4 & 3 & 2 \\ 2 & 3 & 4 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

- In this case for instance ω_3 and ω_4

$$P_1 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3, \omega_4\}\}$$

- The hierarchical index is $f(\{\omega_3, \omega_4\}) = 3$
- The aggregation index of the new cluster is:

$$\delta(\{\omega_1\}, \{\omega_3, \omega_4\}) = \min\{s(\{\omega_1\}, \{\omega_3\}), s(\{\omega_1\}, \{\omega_4\})\} = 1$$
$$\delta(\{\omega_2\}, \{\omega_3, \omega_4\}) = \min\{s(\{\omega_2\}, \{\omega_3\}), s(\{\omega_2\}, \{\omega_4\})\} = 2$$

and we obtain the matrix

$$\begin{pmatrix} - & 1 & 1 \\ 1 & - & 2 \\ 1 & 2 & - \end{pmatrix}$$

3- *We repeat the previous step*

- In this case the maxim value of the aggregation index is obtained with $\{\omega_2\}$ and $\{\omega_3, \omega_4\}$

$$P_2 = \{\{\omega_1\}, \{\omega_2, \omega_3, \omega_4\}\}$$

- The hierarchical index is $f(\{\omega_2, \omega_3, \omega_4\}) = 2$

- The aggregation index of the new classes are calculated:

$$\delta(\{\omega_1\}, \{\omega_2, \omega_3, \omega_4\}) = \min\{\delta(\{\omega_1\}, \{\omega_3, \omega_4\}), \delta(\{\omega_1\}, \{\omega_2\})\} = 1$$

4- *We repeat the step* ***2***

- In this case only the clusters are left to be joined $\{\omega_1\}$ and $\{\omega_2, \omega_3, \omega_4\}$

$$P_3 = \{\{\omega_1, \omega_2, \omega_3, \omega_4\}\} = \Omega$$

- The hierarchical index is $f(\{\omega_1, \omega_2, \omega_3, \omega_4\}) = 1$

Designing a P System

Let $P_{Nk} = (\omega_{ij})_{1 \le i \le N, 1 \le j \le k}$ the value matrix of the individuals to classify:

$$\Pi(P_{Nk}) = (\Gamma(P_{Nk}), \mu(P_{Nk}), M_1, M_2, \ldots, M_N, R, \rho)$$

- Membrane structure: $\mu(P_{Nk}) = [_N [_1]_1 [_2]_2 \ldots [_{N-1}]_{N-1}]_N$

- Working alphabet:

$$\{e_{js}, \quad d_{js} : 1 \le j \le N, 1 \le s \le k\} \bigcup \{a_s, \quad b_s : 1 \le s \le k\} \bigcup$$
$$\{S_{ij}, \quad C_{ij} : 1 \le i < j \le N\} \bigcup$$
$$\{\alpha_{ijt}, X_{ijt} : 1 \le i < j \le N, 1 \le t \le k-1\} \bigcup \{\gamma_i : 1 \le i \le N\} \bigcup$$
$$\{\eta_i : 0 \le i \le (N-1)(3k-1)\} \bigcup \{\epsilon_i : 0 \le i \le 3k-2\} \bigcup$$
$$\{\beta_i : 0 \le i \le k-2\} \bigcup \{\sharp\}$$

- Initial multisets:

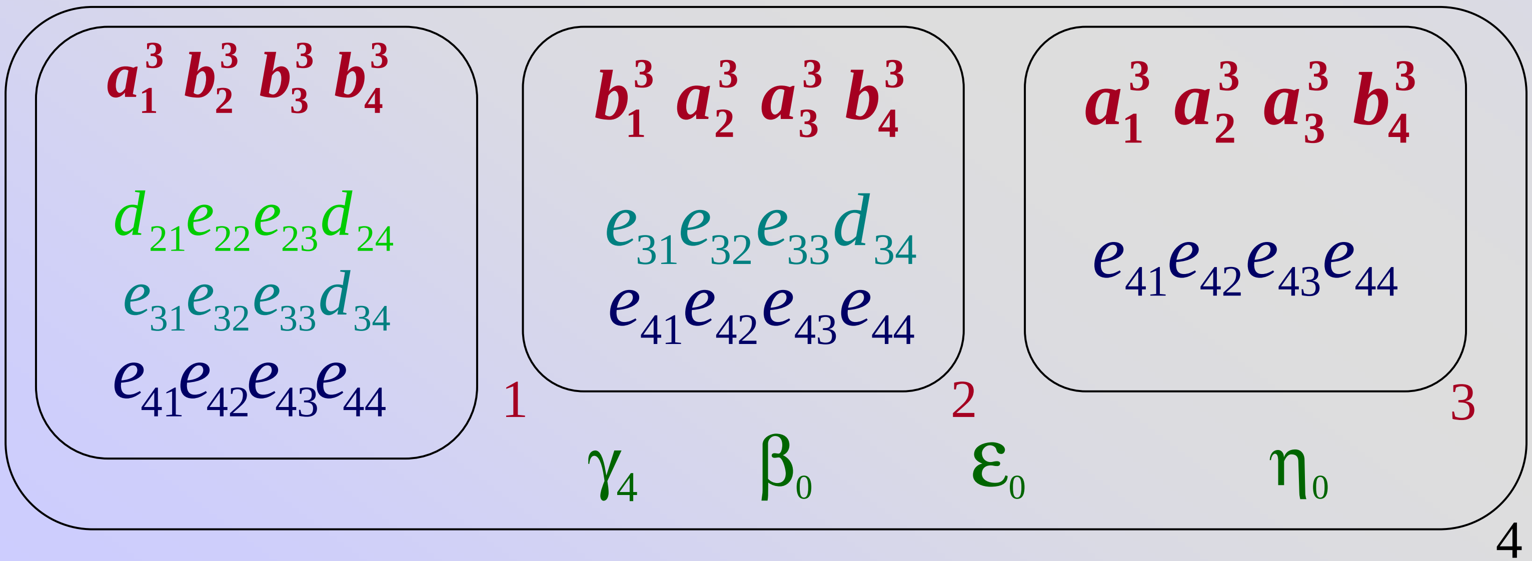
$$\mathcal{M}_i = \{a_s^{(N-i)\omega_{is}} : 1 \le s \le k \wedge 1 \le i \le N-1\} \cup$$
$$\{b_s^{(N-i)(1-\omega_{is})} : 1 \le s \le k \wedge 1 \le i \le N-1\}$$
$$\{e_{js}^{\omega_{js}} : 1 \le s \le k \wedge i \le j \le N\} \cup$$
$$\{d_{js}^{(1-\omega_{js})} : 1 \le s \le k \wedge i \le j \le N\};$$

$$\mathcal{M}_N = \{\gamma_N, \quad \epsilon_0, \quad \eta_0\}$$

An example

$\omega_1 = (1,0,0,0)$ $\omega_2 = (0,1,1,0)$ $\omega_3 = (1,1,1,0)$ $\omega_4 = (1,1,1,1)$

The initial configuration is



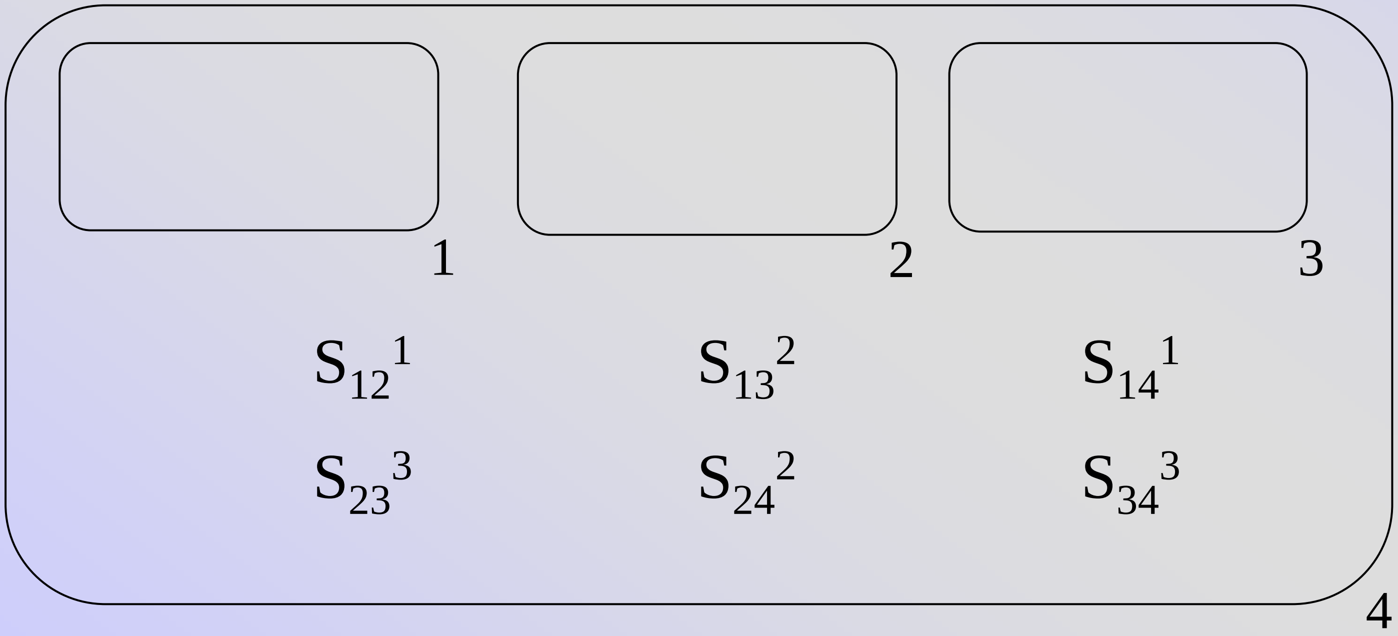
The structure of computation:

1. calculate the similarity between individuals

   - *membrane i* $(1 \le i \le N-1)$

   - 1 step

   - rules: $r_{k+4} = \{a_s e_{js} \to (S_{ij}, out) : 1 \le s \le k, i+1 \le j \le N\}$
     $r_{k+5} = \{b_s d_{js} \to (S_{ij}, out) : 1 \le s \le k, i+1 \le j \le N\}$



2. select the two closest clusters:

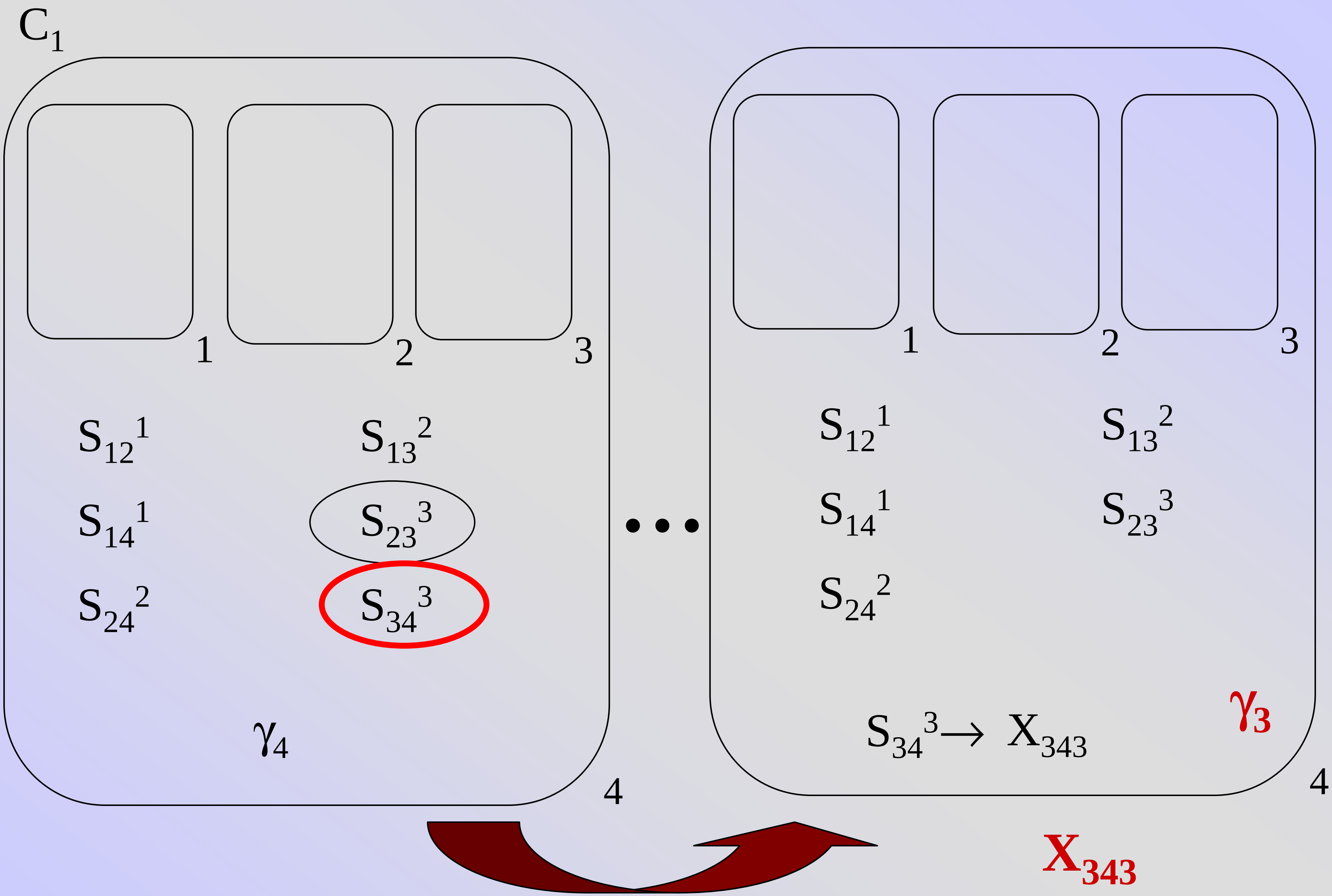
- *skin membrane N*
- k steps
- rules:

$$r_0 = \{\epsilon_0 \rightarrow \epsilon_1\beta_0\} \cup \{\eta_i \rightarrow \eta_{i+1} : 0 \leq i \leq (N-1)(3k-1)-1\}$$
$$\cup \{\epsilon_i \rightarrow \epsilon_{i+1} : 1 \leq i \leq 3k-2 \wedge i \neq k\}$$

$$r_u = \{\beta_{u-1}S_{ij}^{k-u} \rightarrow \alpha_{ij(k-u)} : 1 \leq i < j \leq N\}$$
$$r'_u = \{\beta_{u-1} \rightarrow \beta_u\} \qquad 1 \leq u \leq k-1$$
$$r'_{k-1} = \{\eta_{(N-1)(3k-1)} \rightarrow (\sharp, out)\}$$
$$r_k = \{\epsilon_k\gamma_q\alpha_{ijt} \rightarrow \epsilon_{k+1}X_{ijt}^{q-2}\gamma_{q-1}(X_{ijt}, out) : 2 \leq q \leq N, 1 \leq i < j \leq N,\ 1 \leq t \leq k-1\}$$
$$r'_k = \{\epsilon_k \rightarrow \epsilon_{k+1}\}$$

- priorities:

$$\{r_k > r'_k\} \cup \{r_1 > r'_1 > r_2 > r'_2 > \ldots > r_{k-1} > r'_{k-1}\}$$

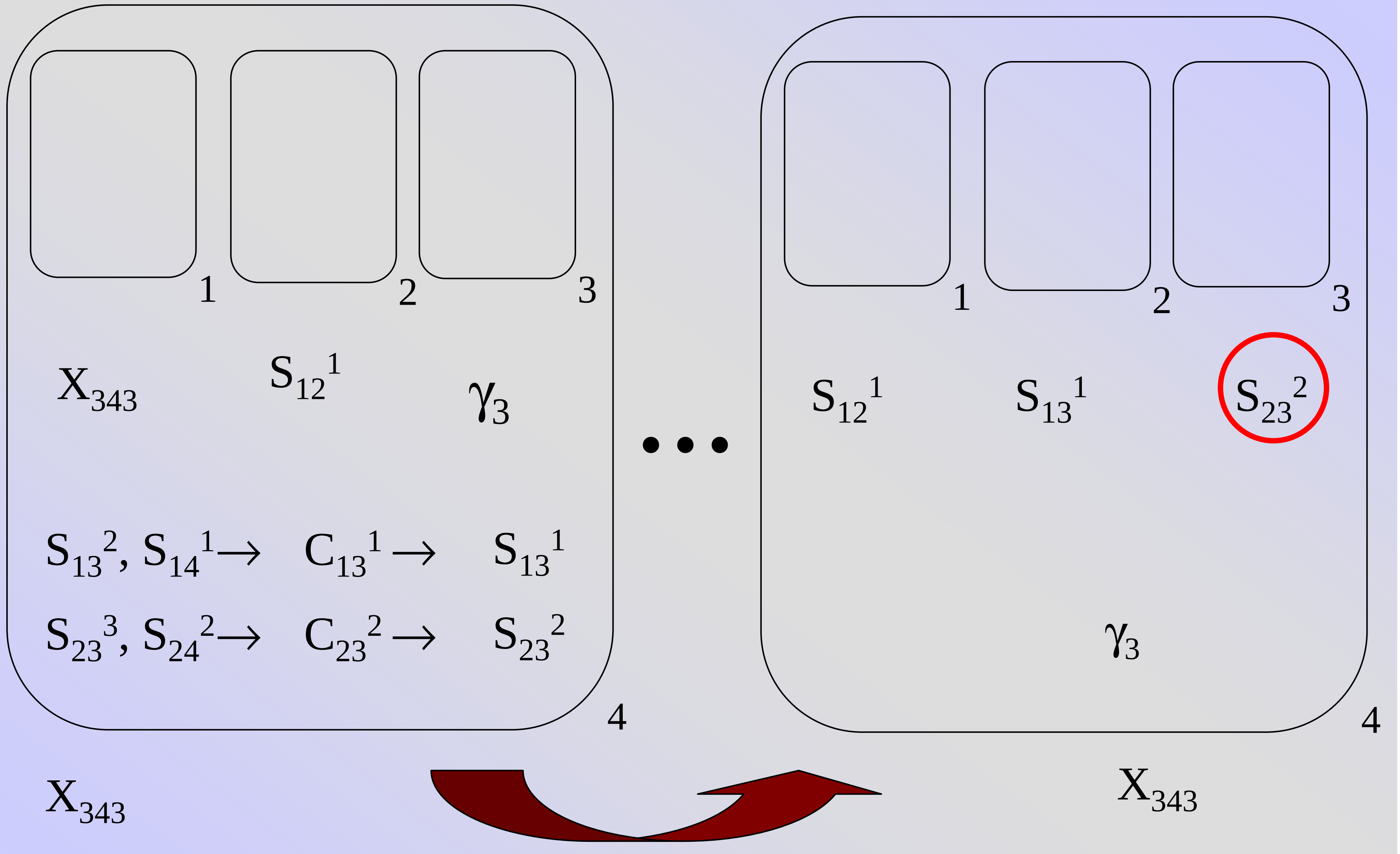

C1
1
2
3
S12 1
S13 2
S14 1
S23 3
S24 2
S34 3
γ4
4
• • •
1
2
3
S12 1
S13 2
S14 1
S23 3
S24 2
S34 3→ X343
γ3
4
X343

3. **compute the aggregation index between new cluster and the other clusters:**

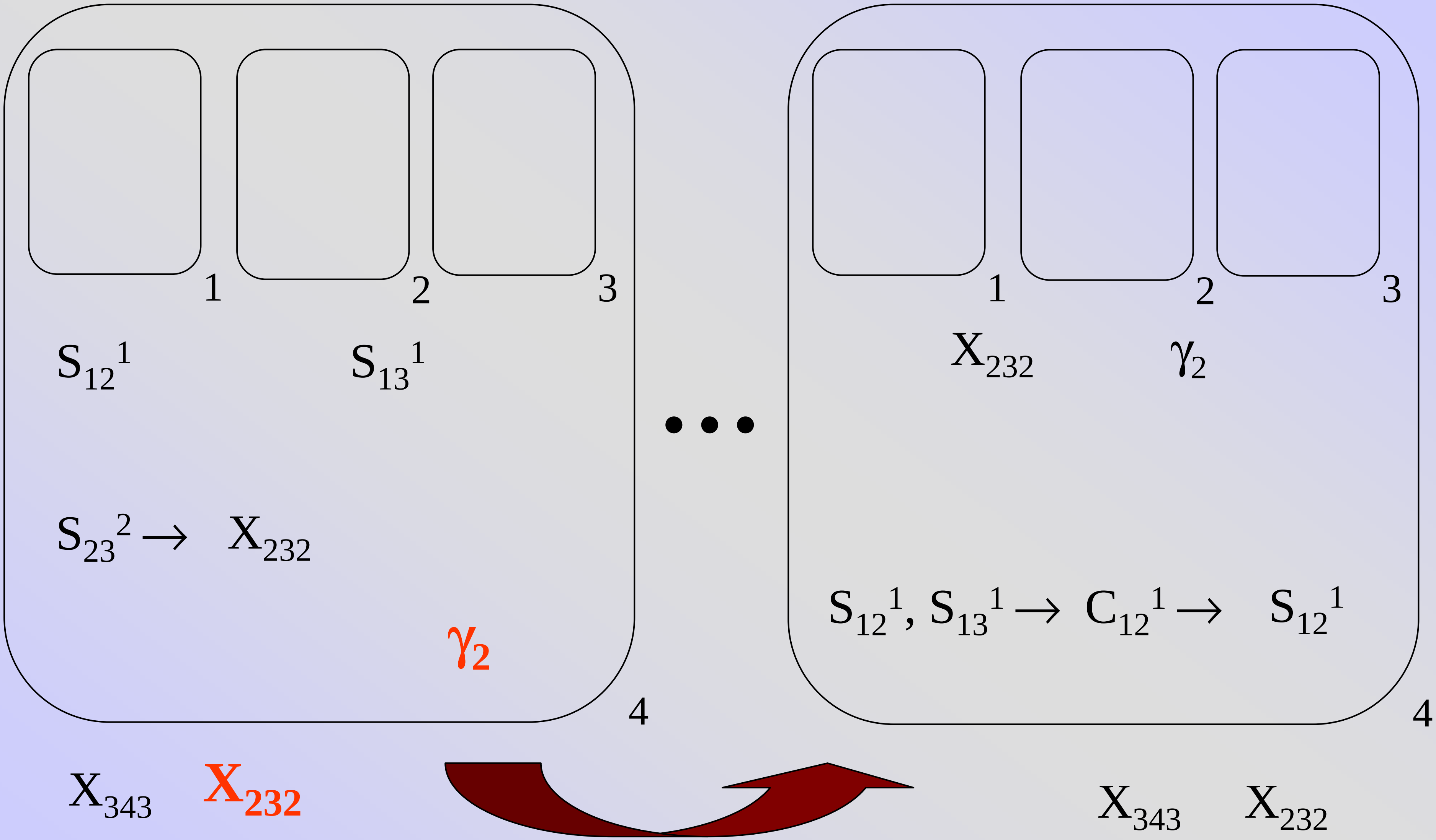
- *skin membrane N*
- 2k-1 steps
- rules:

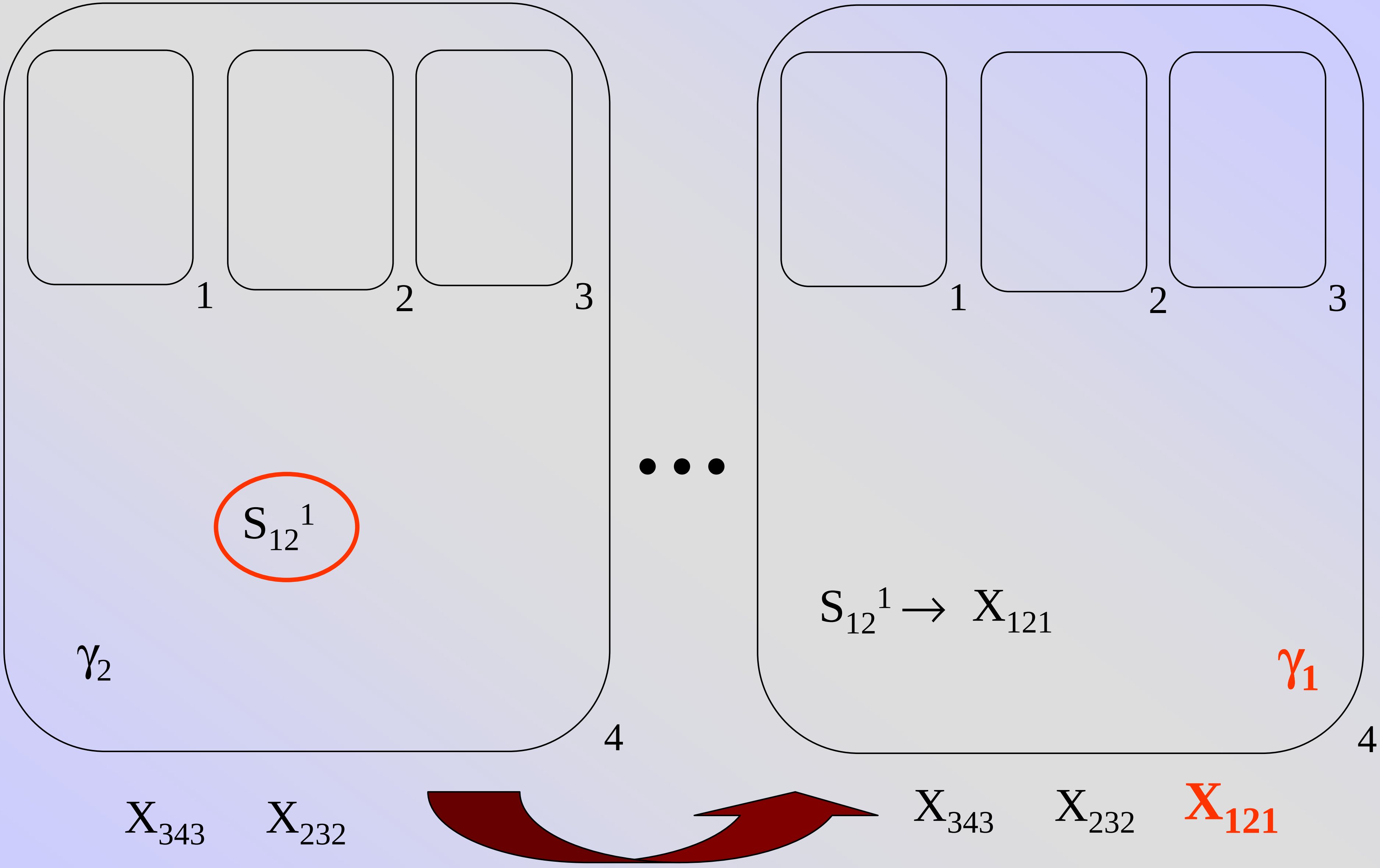
$$r_{k+1} = \{X_{ijt}S_{ip}S_{jp} \rightarrow C_{ip}X_{ijt} : 1 \leq i < j < p \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{ip}S_{pj} \rightarrow C_{ip}X_{ijt} : 1 \leq i < p < j \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{pi}S_{pj} \rightarrow C_{pi}X_{ijt} : 1 \leq p < i < j \leq N,\ 1 \leq t \leq k-1\}$$
$$r_{k+2} = \{X_{ijt}S_{ip} \rightarrow X_{ijt} : 1 \leq i < p < j \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{jp} \rightarrow X_{ijt} : 1 \leq i < j < p \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{pi} \rightarrow X_{ijt} : 1 \leq p < i < j \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{pj} \rightarrow X_{ijt} : 1 \leq p < i < j \leq N,\ 1 \leq t \leq k-1\}$$
$$\{X_{ijt}S_{pj} \rightarrow X_{ijt} : 1 \leq i < p < j \leq N,\ 1 \leq t \leq k-1\}$$
$$r_{k+3} = \{C_{ij} \rightarrow S_{ij} : 1 \leq i < j \leq N\}$$
$$\{\epsilon_{3k-1}X_{ijt}^{q-2}\gamma_{q-1} \rightarrow \epsilon_1\beta_0\gamma_{q-1} : 1 \leq i < j \leq N,\ 1 \leq t \leq k-1\}$$

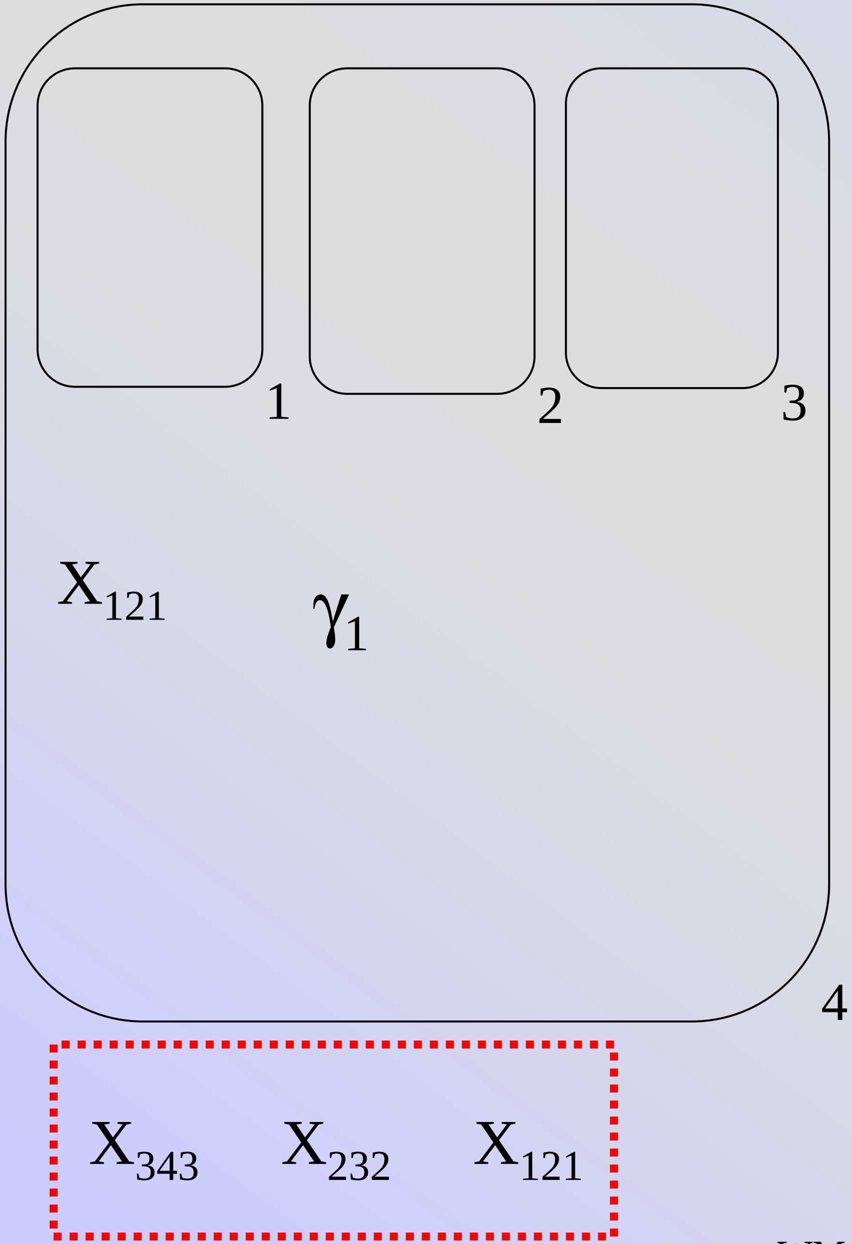
- priorities: $\{r_{k+1} > r_{k+2} > r_{k+3} > r'_{k+3}\}$


1
2
3
X_{343}
S_{12}^{1}
γ_3
$S_{13}^{2}, S_{14}^{1} \rightarrow C_{13}^{1} \rightarrow S_{13}^{1}$
$S_{23}^{3}, S_{24}^{2} \rightarrow C_{23}^{2} \rightarrow S_{23}^{2}$
4
X_{343}
• • •
1
2
3
S_{12}^{1}
S_{13}^{1}
S_{23}^{2}
γ_3
4
X_{343}

4. repeat the steps 2 and 3 *N-1* times:




1
2
3
S_{12}^{1}
γ_2
4
1
2
3
$S_{12}^{1} \rightarrow X_{121}$
γ_1
4
X_{343} X_{232}
X_{343} X_{232} X_{121}



- Initially: $h_1 = \{\omega_1\} \quad h_2 = \{\omega_2\}$

  $h_3 = \{\omega_3\} \quad h_4 = \{\omega_4\}$

- From the object X_{343} we obtain:

  $h_1 = \{\omega_1\} \quad h_2 = \{\omega_2\}$

  $h_3 = \{\omega_3, \omega_4\}$

- From the object X_{232} we obtain:

$$h_1 = \{\omega_1\} \quad h_2 = \{\omega_2, \omega_3, \omega_4\}$$

- Finally from the object X_{121}

  we obtain:

  $h_1 = \{\omega_1, \omega_2, \omega_3, \omega_4\}$

- **The resources initially required for constructing the P system $\Pi(P_{Nk})$ from the matrix P_{Nk} are:**

  - Initial number of membranes: N
  - Size of the alphabet: $\theta(k \cdot N^2)$
  - Sum of the sizes of initial multisets: $\theta(k \cdot N)$
  - Number of rules: $\theta(k \cdot N^3)$
  - Maximum of rules' lengths: $\theta(N)$
  - Number of priority relations: $\theta(k^2 \cdot N^6)$
  - Cost of time: $\theta(k \cdot N)$

- **In classical algorithm:**

  - Number of steps: $\theta(k \cdot N^3)$

Formal Verification

For each matrix P_{Nk} we construct a P system $\Pi\left(P_{Nk}\right)$. We must to prove that:

- The multiplicity of the objects $S_{ij} \in C_1\left(N\right) \quad \rightarrow t_{ij}^{(1)}$

$$t_{ij}^{(1)} = s\left(\omega_i, \omega_j\right)$$

- The multiplicity of the objects $S_{ij} \in C_{1+n(3k-1)}\left(N\right) \quad \rightarrow t_{ij}^{(n)}$

$$t_{ij}^{(n)} = \delta\left(B_i^{(n)}, B_j^{(n)}\right) \quad B_i^{(n)}, B_j^{(n)} \in \Delta_n$$

- The halting configuration is $C_{(N-1)(3k-1)}$

- The objects $X_{i_n j_n t_{i_n j_n}^{(n)}}$ are sent to environment only in $C_{1+k+n(3k-1)}$

and $t_{i_n j_n}^{(n)} = \max\{t_{ij}^{(n)} \mid B_i^{(n)}, B_j^{(n)} \in \Delta_n\}$

$$B_{i_n}^{(n+1)} = B_{i_n}^{(n)} \cup B_{j_n}^{(n)} \in \Delta_{n+1}$$

- With the partition $\Delta_0, \Delta_1, \ldots, \Delta_{N-1}$ we can construct a hierarchy

Conclusions

- A P system associated with a set of individuals is presented.

- This P system provides one possible hierarchy.

- The solution is encoded in the environment of the P system

- The solution provide is:

  - **Semi-uniform**

  - **Efficient**

- The amount of resources initially required to construct the P system is quadratic in N and k.

- ***ADVANTATGE:***

✓ The cost of time of the P system is quadratic in *N* and *k*.

✓ In classical algorithm the number of steps:

$$\theta(k \cdot N^{3})$$