Analyzing a Database of Digitalized Process Flowsheets to Extract Relevant Initialization Structures for Evolutionary Process Synthesis Methods

Jean-Marc Commenge\*, Andres Pina-Martinez

*Université de Lorraine, CNRS, LRGP, F-54000 Nancy, France*

*jean-marc.commenge@univ-lorraine.fr*

Abstract

Process synthesis using evolutionary methods, based on the iterative application of mutation operators, requires to initialize the method by one or a set of process structures. Appropriate initialization might reduce computation times by providing first proposals that reduce the number of mutations to reach optimal structures, in terms of units and connectivity. This work demonstrates how to identify, from a given database of flowsheets, the most-central flowsheets that might play a pivotal role in the evolutionary approach. Three centrality criteria (outcloseness, betweenness and pagerank) are combined in a Pareto analysis to ensure centrality and diversity of the initial population.

**Keywords**: process synthesis, flowsheet, database, evolutionary distance, centrality.

**1. Introduction**

The energy transition requires certain chemical processes to be redesigned to adapt to sustainable operation (biobased inputs, intermittency, decentralized production, etc.) and new processes to be created from scratch, increasing the need for fast, or even automatic, process synthesis methods. These methods have considerably evolved from heuristics-based to mathematical-programming based optimization (leading often to MINLP problems): by reducing the influence of prior knowledge and expert bias, evolutionary approaches enlarge the scope of attainable structures, but may be resource demanding to converge to optimal structures. Evolutionary approaches differ in the way they integrate the evolutionary feature, either by including a genetic algorithm in the optimization of a proposed population or superstructure (Koch et al., 2007; Skiborowski et al., 2015), or by applying mutation operators to evolving process structures (Wang et al., 2015; Neveux, 2018). In all cases, proper initialization with unbiased complex process structures should enable accelerate convergence while enlarging the explored domain.

To identify appropriate initialization structures, a database of digitalized process flowsheets has been created, which enables to visualize statistical distributions of these processes with respect to their number of units and streams, that directly relates to process engineering strategies such as process integration and intensification. Analysis of mutations operators enables to define a semi-distance between structures: the flowsheet distribution can then be studied according to centrality criteria. Three centrality criteria are computed (outcloseness, betweenness and pagerank) yielding to a Pareto front ensuring diversity and relevance of selected structures. This work presents the database of digitalized flowsheets, the main features of their statistical distributions, the computation of the evolutionary distance and the most central process flowsheets with respect to the centrality criteria.

**2. Database of digitalized flowsheets**

*2.1. General features*

The database used for this analysis contains 767 flowsheets, digitalized from 278 papers published since 2006. They include 9971 units (including process inputs and outputs) and 10721 streams. For each flowsheet, the database provides an entry name, the number of streams, unit operations, process inputs, outputs and units among a list of 23 types (Absorber, Compressor, Cyclone, Distillation, etc.). The ordered list of units, with meaningful prefixes and numbers to distinguish duplicate units, is provided as well as the incidence matrix: each column corresponds to a unit, and each line to a stream with values -1 or +1 to identify upstream or downstream units respectively. Additional data include the flowsheet image, keywords, species and the scientific reference. Among the 9971 units, all 23 types of units are present despite a large domination of 1263 heat exchangers, 1173 distillation columns and 818 mixers compared to only 4 filters, 5 dryers and 9 mills.

*2.2. Statistical distribution of the flowsheet’s population*

Figure 1 presents projections of the population as a function of the number of process inputs and outputs (left) and number of streams and unit operations (right). Grey-shaded areas indicate impossible zones: except 6 cycles which possess no input or output, a process with a non-null number of inputs (resp. outputs) must contain a non-null number of outputs (resp. inputs). Also, a process with a number of units larger than the number of streams plus one is not feasible. In Figure 1 (left), despite a mode corresponding to 2 inputs and 2 outputs with 131 structures, the distribution is asymmetric with a medoid corresponding to 2 inputs and 3 outputs. This large number of process outputs with respect to the inputs might result from the vast literature dedicated to separation processes. Despite the large number of inputs/outputs of some flowsheets, up to a total of 16, almost 90 % of them exhibit less than 4 inputs and 4 outputs.



Figure 1: Projections of the distribution of digitalized flowsheets as a function of the number of process inputs and outputs (left) and number of streams and unit operations (right).

In Figure 1 (right), the distribution ranges from a single stream (Neveux et al., 2018) up to a complex power-to-liquids process with 55 units and 50 streams (Gao et al., 2022), with a dense core around 10 units and 10 streams. The distribution shape is governed by the average degree of the process graphs. In a graph representing a process, the degree of a vertex (process unit) is the number of edges (process streams) adjacent to this vertex.

The average degrees of the processes range from 1 (single stream) up to 3.28, with a global average of 2.16. Only a few dozen processes exhibit an average degree above 2.5, whereas the majority lies between 2 and 2.5. Despite these low average degrees, some processes contain highly-connected units, mainly multi-fluid heat exchangers, with an individual degree up to 12.

**3. Evolutionary methods, mutations and evolutionary distance**

*3.1. Evolutionary methods, mutations and their impact on process structures*

Evolutionary methods based on step-wise modifications of process structures by application of mutation operators mainly consider insertion and removal of units, and streams permutations (Wang et al., 2015; Neveux, 2018). In the present work, these operators are parameterized with respect to the degree of the inserted/removed units, and a new operator consisting in merging one input and one output is added (Table 1). For each operator, Table 1 indicates its impact on the total number of inputs and outputs, the number of streams and number of units. For example, inserting a distillation column of degree 3 (D = 3) increases the number of inputs and outputs by one unit, whereas the numbers of streams and units increase by two.

Table 1: Impact of the application of various mutation operators on the total number of process inputs and outputs, number of streams and number of units. D denotes the degree of a unit.

|  |  |  |  |
| --- | --- | --- | --- |
| Mutation operator | Ninputs + Noutputs | Nstreams | Nunits |
| Permute stream connections | = | = | = |
| Add a Dth-degree unit $(D\geq 2)$ | $$+D-2$$ | $$+D-1$$ | $$+D-1$$ |
| Remove a Dth-degree unit $(D\geq 2)$ | $$+D-2$$ | $$-1$$ | $$+D-3$$ |
| Merge one input and one output | $$-2$$ | $$-1$$ | $$-2$$ |

*3.2. Definition and computation of an evolutionary (semi-)distance*

The evolutionary distance used to select the most appropriate process structures for initialization is defined as the minimum number of mutations to be performed to transform a given process structure Psource into another given process structure Ptarget, under a set of authorized mutations, excluding stream permutations. Being given two structures Psource and Ptarget with their lists of units and incidence matrices, the computation of the distance d(Psource, Ptarget) includes the following steps:

1. List all units (except inputs and outputs) and corresponding degrees in the source and target structures,
2. Compare the number of inputs and outputs and number of streams of structures to calculate the required increments Inputs + Outputs,req and streams,req,
3. Identify all units (with type and degree) that are not common to both structures,
4. Initialize actual increments Inputs+Outputs and streams and d(Psource,Ptarget) to 0,
5. Update actual increments Inputs + Outputs and streams according to Table 1, and increase the distance d(Psource, Ptarget) by one unit for all necessary mutations:
	1. Remove all not-common units included in the source structure,
	2. Add all not-common units missing in the target structure,
6. Merge one input and one output of the target structure as many times as required (and accordingly increase the distance d(Psource, Ptarget) by one unit) so that Inputs+Outputs and streams be equal to Inputs + Outputs,req and streams,req.

This computation sequence provides a measurement of the difference between structures, more specifically a semi-distance: indeed, whereas identity and positivity conditions are obvious, the symmetry condition is not satisfied: d(Psource, Ptarget) is not always equal to d(Ptarget, Psource). The triangle inequality has been checked on 60 million triplets (P1, P2, P3). This lack of symmetry slows down the computation of the distance matrix and hinders some post-treatment methods. When necessary, a real distance can be defined by averaging d(Ptarget, Psource) and d(Psource, Ptarget).

Figure 2 (left) presents the distribution of evolutionary distances between all couples of process structures in the database. Whereas 480 couples exhibit a null distance, indicating the presence of similar processes in the database that only differ by their connections, the average distance is about 17.5 mutations, confirming the diversity of the flowsheets. The maximum distance is as large as 70, demonstrating that even large flowsheets can include radically different units in terms of type or connectivity. The shape of the distribution does not enable to identify any significant structure (clusters, gaps, etc.) in the population.

|  |  |
| --- | --- |
|  |  |

Figure 2: Distribution of evolutionary distances (left) and visualization of the population of flowsheets by 2D multi-dimensional scaling (right).

A 2D multidimensional scaling (MDS) is presented in Figure 2 (right) (Aggarwal, 2015). MDS is a projective method for dimensionality reduction that optimizes a distribution of points (here on a plane) that exhibit a distance matrix similar to the original data. Despite its projective character, this view confirms the lack of structure in the population, that looks like a gaussian distribution. The maximum distance between the points on the far right and those on the far left, about 70, is coherent with the distance distribution.

**4. Analysis of most-central flowsheets**

The population of flowsheets can be considered as a graph whose nodes are the process structures and whose edges are weighted by the distance between structures. To identify flowsheets that might be relevant for initialization of evolutionary approaches, three centrality criteria are computed for the database flowsheets (Aggarwal, 2015):

* Outcloseness measures how many mutations are required to access any other structure j from a structure i. Centrality is maximal when the sum of distances is minimal, explaining the use of the inverse of the sum of distances as:

|  |  |
| --- | --- |
| $$J\_{outcloseness}\left(i\right)=\frac{1}{\sum\_{j\ne i}^{}d(i,j)}$$ | (1) |

* Betweenness measures how often each process structure appears on a shortest path between two other structures in the population. For structures s and t, the total number of shortest paths between them is denoted Nst, and the number of shortest paths including the structure i is denoted nst(i):

|  |  |
| --- | --- |
| $$J\_{betweenness}\left(i\right)=\sum\_{s,t\ne i}^{}\frac{n\_{st}(i)}{N\_{st}}$$ | (2) |

* Pagerank measures the probability of presence at the node i of the graph of flowsheets during a random walk through this graph. At each node in the graph, the next node is chosen from all downstream nodes with a probability inversely proportional to their evolutionary distance. The pagerank can be interpreted as the average time spent at each node during this random walk.

Figure 3 presents the centrality criteria for all structures as a function of their numbers of streams and units. Structures of interest possess maximal values of the criteria. All criteria exhibit similar trends, with a decrease as a function of the number of streams and units. Betweenness is the most discriminatory whereas outcloseness and pagerank have similar shapes. Large structures mainly exhibit low centrality but exceptions exist: some large processes with up to 48 units exhibit medium centrality. According to outcloseness and betweenness, best structures are small including less than 5 streams/units, whereas pagerank favors structures with 6 to 10 streams/units, indicating that a compromise might be necessary to ensure relevance and diversity of the initialization.

|  |  |  |
| --- | --- | --- |
| Outcloseness | Betweenness | Pagerank |
|  |  |  |

Figure 3: Evolutions of the three centrality criteria for all structures in the database as a function of their numbers of streams and unit operations.

To consider all criteria, a multi-objective analysis is performed in the 3D space composed by the criteria. Projections in Figure 4 highlight with circles the 20 non-dominated structures according to the three criteria. The similar trends observed in Figure 3 reduce the number of Pareto structures while maintaining diversity of their central character. Only 8 circles (with overlap) are visible in Figure 4 confirming the presence of similar structures (subject to stream permutations).

Figure 5 presents the process flowsheets of the 8 different types observed along the Pareto front. Except inputs and outputs, they only contain distillation columns, mixers and splitters. The simplest one is the single stream (most central structure with respect to single-objective optimization of outcloseness), and the most complex includes three columns with a recycle stream (most central structure with respect to single-objective optimization of pagerank).

Despite the surprising absence of heat exchangers in the optimal solutions (most frequent unit type in the database), the dominating presence of distillation columns in the database might be a bias that favors this unit type: reduction of the database should be considered on a case by case basis to fit to case studies specifications.



Figure 4: Projections of the database structures over the planes of the three-dimensional space of the three criteria. Circles indicate the structures on the Pareto front.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 2 variants | 8 variants | 4 variants | 2 variants |

Figure 5: Visualization of the 8 different types of process structures on the Pareto front.

**5. Conclusions**

Analysis of a flowsheets database enabled to identity 8 types of process structures (and their variants by stream permutation) that are the most central with respect to three centrality criteria (outcloseness, betweenness and pagerank). This pivotal position among the population of flowsheets makes them relevant for initialization of evolutionary approaches for process synthesis, while ensuring some diversity of the initialization. Effective impact of this initialization on the complete process synthesis computation still needs to be quantified as well as the possible introduction of biases.

**References**

C. Aggarwal, 2015, Data Mining – the textbook, Springer.

R. Gao, L. Zhang, L. Wang, X. Zhang, C. Zhang, K.-W. Jun, S. K. Kim, H.-G. Park, Y. Gao, Y. Zhu, T. Zhao, H. Wan, G. Guan, 2022, A comparative study on hybrid power-to-liquids/power-to-gas processes coupled with different water electrolysis technologies, Energy Conversion and Management, 263, 115671.

C. Koch, F. Cziesla, G. Tsatsatonis, 2007, Optimization of combined cycle power plants using evolutionary algorithms, Chemical Engineering and Processing, 46, 1151-1159.

T. Neveux, 2018, Ab-initio process synthesis using evolutionary programming, Chemical Engineering Science, 185, 209-221.

M. Skiborowski, M. Rautenberg, W. Marquardt, 2015, A hybrid evolutionary-deterministic optimization approach for conceptual design, Ind. Eng. Chem. Res., 54, 10054-10072.

L. Wang, P. Voll, M. Lampe, Y. Yang, A. Bardow, 2015, Superstructure-free synthesis and optimization of thermal power plants, Energy, 91, 700-711.