Searching for substructures in polyhedral topologies of inorganic crystal structures

Hans-Joachim Klein
Institut f. Informatik u. Prakt. Mathematik, Technische Fakultät
Christian-Albrechts-Universität Kiel, Germany

www.is.informatik.uni-kiel.de/~hjk/crystana.html

Abstract

When large collections of crystal structures are studied at the level of coordination polyhedra, advanced facilities for searching related figures formed by polyhedra are of interest. Well-known indexing techniques based upon predefined sets of rigid substructures such as rings, for instance, are in general difficult to apply in connection with inorganic crystal structures because of the large variety of chemical elements and patterns which have to be taken into account. Hence we have built a system based upon a representation of polyhedral topologies by graphs which uses an indexing technique not referring to special substructures but to paths in graphs. For any finite cluster of polyhedra it allows to find all possible embeddings into polyhedral topologies of a given set of crystal structures.
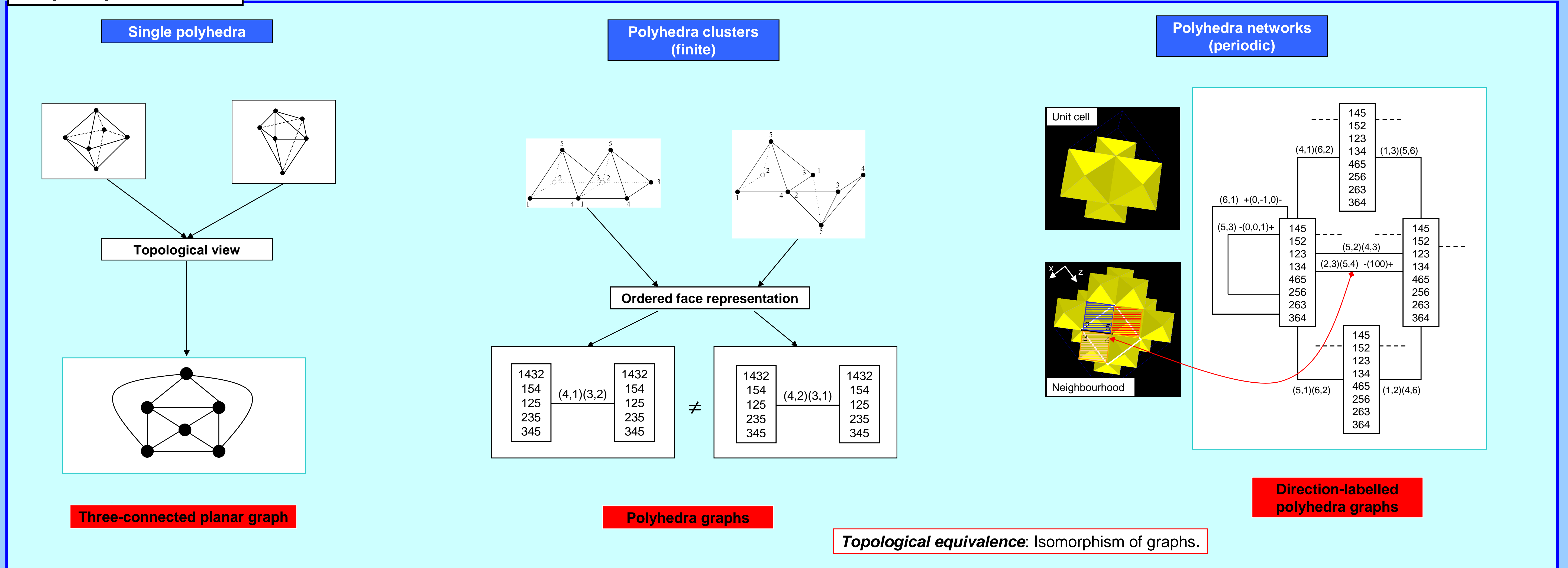
In *polyhedra graphs*, a coordination polyhedron is represented as a single node and a connection between two polyhedra as a labelled line. With each node an ordered face representation of the corresponding polyhedron is associated. For every pair of neighbouring polyhedra sharing a vertex, an edge, or a face the corresponding nodes of the graph are connected by a line which is labelled with the pairs of numbers of polyhedral vertices involved in the connection. Finite graph representations can be obtained for infinite topologies by folding infinite polyhedra graphs appropriately. Depending on the intended usage of the graph representation, additional information such as symmetries or coordinates can be assigned to the nodes.
Let a finite input cluster C of polyhedra and a polyhedra graph P of a crystal structure be given. C is *topologically embeddable* into P if the polyhedra graph of C is isomorphic to a subgraph of P containing with each pair of nodes all lines between these nodes in P.
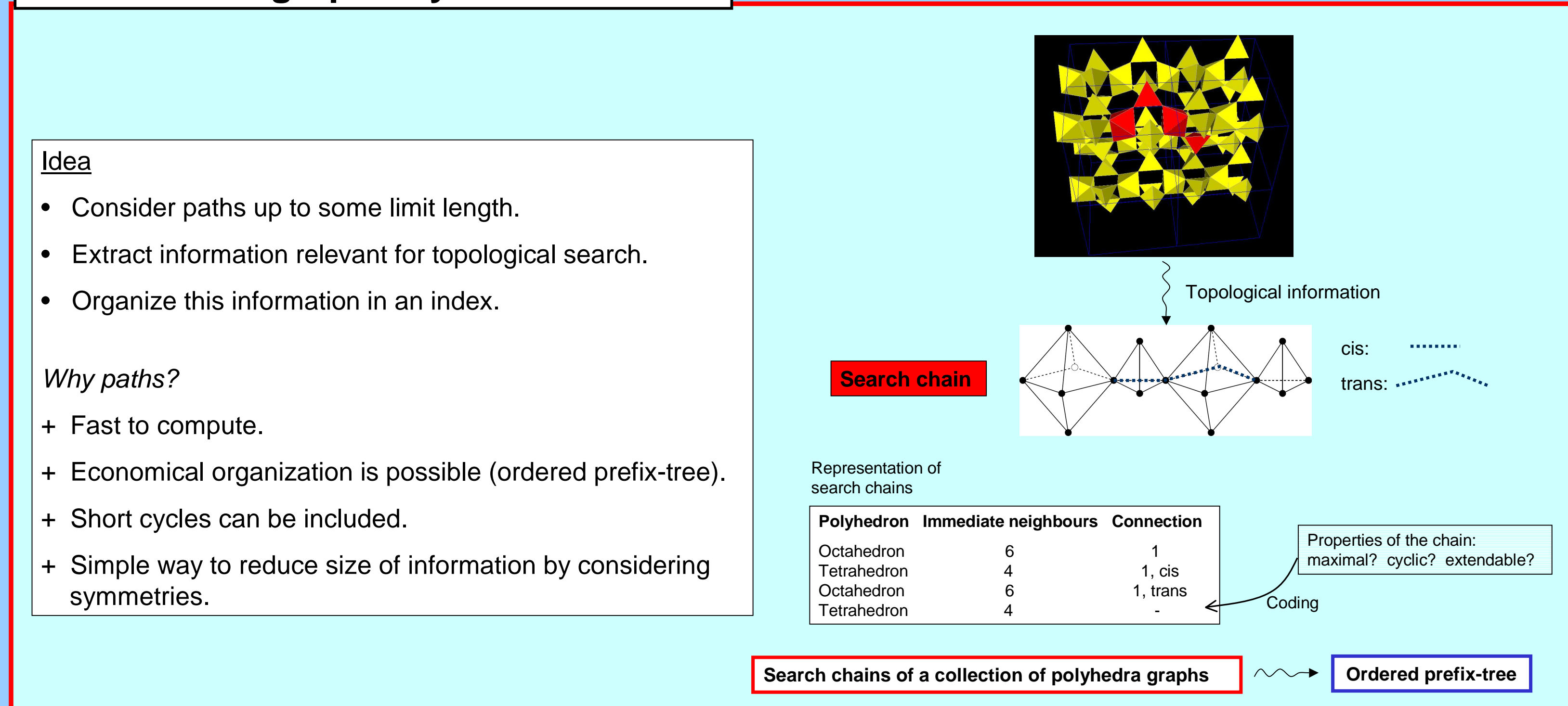
To deal with the computational complexity of the subgraph isomorphism problem, the structures of a given collection are preprocessed and an index is built based upon a path enumeration for the polyhedra graphs of all structures. Using prefix-trees and a fixed interval for the lengths of paths to consider, a space-saving indexation can be achieved. For the graph of the input cluster a minimal line covering by paths is determined and applied for search. This technique allows to search for substructures efficiently even in large collections of data.
A web interface has been realized such that the searching functionality can be accessed on-line. It is possible to mark polyhedra in a three-dimensional interactive display of a structure and to search for one or all possible topological embeddings in a given database. New structures can be added by submitting structure data as a HTML form or as a CIF file.
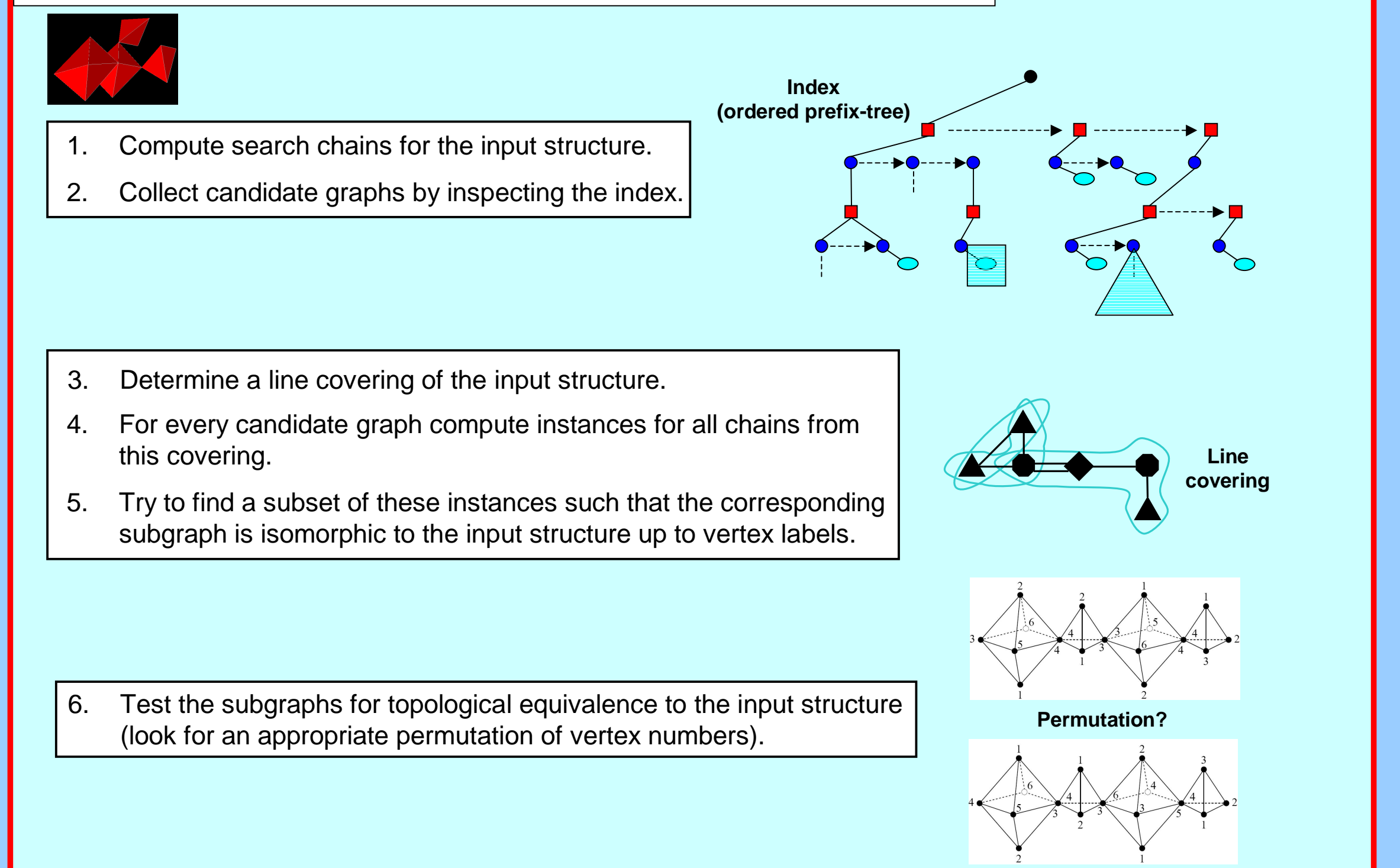
Graph representation

Single polyhedra

Topological view

Three-connected planar graph

Polyhedra clusters (finite)

Ordered face representation

1432		1432
154	(4,1)(3,2)	154
125		125
235		235
345		345

$\neq$

1432		1432
154	(4,2)(3,1)	154
125		125
235		235
345		345

Polyhedra graphs

Polyhedra networks (periodic)

Unit cell

Neighbourhood

Direction-labelled polyhedra graphs

Topological equivalence: Isomorphism of graphs.

Indexation of graphs by search chains

Idea
- Consider paths up to some limit length.
- Extract information relevant for topological search.
- Organize this information in an index.

Why paths?
+ Fast to compute.
+ Economical organization is possible (ordered prefix-tree).
+ Short cycles can be included.
+ Simple way to reduce size of information by considering symmetries.

Topological information

Search chain

cis: ......
trans: -- --

Representation of search chains

Polyhedron	Immediate neighbours	Connection
Octahedron	6	1
Tetrahedron	4	1, cis
Octahedron	6	1, trans
Tetrahedron	4	-

Properties of the chain: maximal? cyclic? extendable?

Coding

Search chains of a collection of polyhedra graphs ⟿ Ordered prefix-tree

Determination of isomorphic substructures

1. Compute search chains for the input structure.
2. Collect candidate graphs by inspecting the index.

Index (ordered prefix-tree)

3. Determine a line covering of the input structure.
4. For every candidate graph compute instances for all chains from this covering.
5. Try to find a subset of these instances such that the corresponding subgraph is isomorphic to the input structure up to vertex labels.

Line covering

6. Test the subgraphs for topological equivalence to the input structure (look for an appropriate permutation of vertex numbers).

Permutation?

Interactive graphical interface

"Query"

Database of crystal structures

Result