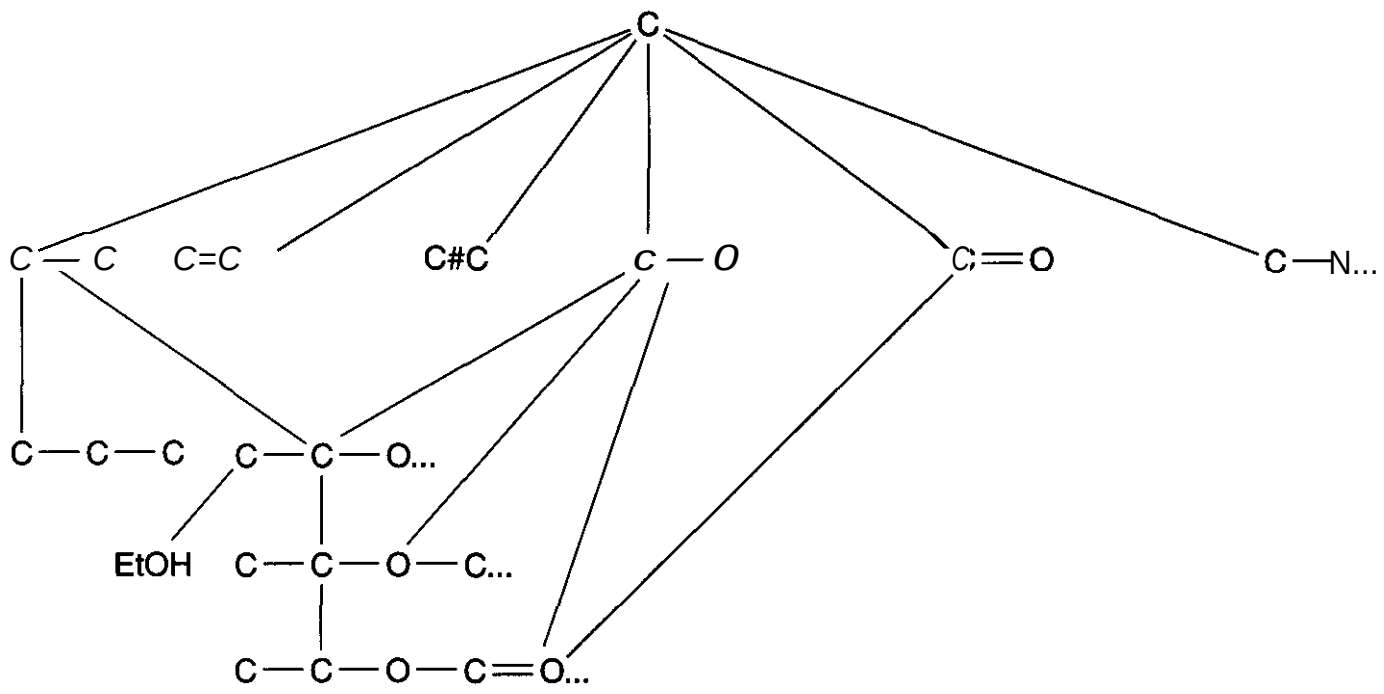

Figure **9.10.** Simplified ISIS Fastsearch index—ethanol is a leaf node that can be reached from several substructure nodes.

ments found in structures in the database, up to a fixed pathlength. These are stored in a highly compressed binary format (Fig. 9.10). Similar approaches have appeared in the literature (64). Leaf nodes in the tree contain identifiers of specific structures in the database (simplified in Fig. 9.10). An exact match or substructure search consists of traversing the tree to find structures in the database that have substructural fragments in common with the query structure. Because the fastsearch index is large—often as large as the rest of the structure database, updating it for the addition or removal of structures is time consuming.

This relational chemical database format is extended in ISIS to include 3D models, generic structures, and most recently, reactions. In these cases, additional "trees" in the database hierarchy connect 2D structures with 3D models, connect root structures with corresponding Rgroup members, or connect molecules with reactions.

Other relational *structure/reaction* database systems are available commercially. These include the Thor system from Daylight (65), Accord and RS^3 Discovery from Accelrys (66), and Unity from Tripos (67). Personal database systems that can be implemented on a desktop computer include ISIS/Base (68), Accord for Access (66), and Team Works from Afferent (69).

3.2 Registering Chemical Information

Chemical structure registration is an important activity that is necessary for drug discovery. The structures that have been developed by a pharmaceutical company constitute the "crown jewels" of chemical information, and they must be properly and securely archived. The registration process usually involves the process of extracting, cleaning, transforming, and loading the data—sometimes termed ECTL.

3.2.1 Extract the Data. First, the *structures/reactions* and corresponding data are extracted, collected, and validated. Increasingly, this is managed automatically, using output from the high-throughput chemistry process. Laboratory information management systems (LIMS) that are "structure smart" can manage chemical structure information starting from the design of a reaction, through the synthesis of the compounds, the chemical analysis of the structures, the in vitro biological assay, and finally the storage in the chemical database. Certain steps, such as drawing the initial structures/reactions, still remain an activity for the chemist, although many chemical information systems can take a generic structure, enumerate the many specific combinations, and layout the structures automatically (for example, the Monomer Toolkit by Day-