

# COMS 4115 – Programming Languages and Translators

---


## LANAGAGUE PROPOSAL

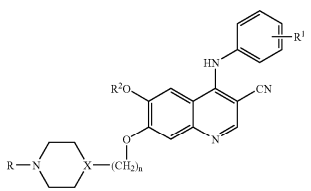
NING YU  
NY2186@COLUMBIA.EDU

# MARKUSH DESCRIPTION LANGUAGE FOR CHEMICAL PATENTS

## I. Background

Close to 30 percent of all patents are chemical patents, of which roughly 10 percent claim IP rights over novel chemical structures using the Markush notation[1-5]. An example of an approved patent with claims on a series of 4-anilino-3-quinolinecarbonitriles-based chemical structures for the treatment of chronic myelogenous Leukemia (CML) is shown below in Figure 1.

  
US007417148B2

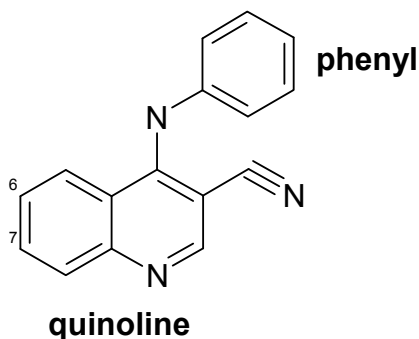
<p>(12) <b>United States Patent</b> <b>Boschelli et al.</b></p> <p>(54) <b>4-ANILINO-3-QUINOLINECARBONITRILES FOR THE TREATMENT OF CHRONIC MYELOGENOUS LEUKEMIA (CML)</b></p> <p>(75) Inventors: <b>Frank Boschelli</b>, New City, NY (US); <b>Kim T. Arndt</b>, Towaco, NJ (US); <b>Jennifer M. Golas</b>, Hewitt, NJ (US)</p> <p>(73) Assignee: <b>Wyeth</b>, Madison, NJ (US)</p> <p>(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 446 days.</p> <p>(21) Appl. No.: <b>10/980,097</b></p> <p>(22) Filed: <b>Nov. 3, 2004</b></p> <p>(65) <b>Prior Publication Data</b> US 2005/0101780 A1 May 12, 2005</p> <p><b>Related U.S. Application Data</b></p> <p>(60) Provisional application No. 60/517,819, filed on Nov. 6, 2003.</p> <p>(51) <b>Int. Cl.</b> <i>C07D 215/38</i> (2006.01) <i>A61K 31/47</i> (2006.01)</p> <p>(52) <b>U.S. Cl.</b> ..... <b>546/159</b>; 546/157; 546/153; 514/313; 514/253.06</p> <p>(58) <b>Field of Classification Search</b> ..... 514/313, 514/253.06; 546/157, 159, 153, 53 See application file for complete search history.</p> <p>(56) <b>References Cited</b></p> <p style="text-align: center;">U.S. PATENT DOCUMENTS</p> <p>6,002,008 A 12/1999 Wissner et al. 6,780,996 B2 8/2004 Boschelli et al.</p> <p style="text-align: center;">FOREIGN PATENT DOCUMENTS</p> <p>WO 03/093241 A1 11/2003 WO WO 2004/075898 A1 9/2004</p> <p style="text-align: center;">OTHER PUBLICATIONS</p> <p>Boschelli, J Med Chem, 2001, vol. 44, pp. 822-833.* Boschelli, J Med Chem, 2001, vol. 44 pp. 3965-3977.* Registry compund No. 220127-57-1, Mar. 3, 1999.*</p>	<p>(10) <b>Patent No.:</b> <b>US 7,417,148 B2</b></p> <p>(45) <b>Date of Patent:</b> <b>Aug. 26, 2008</b></p> <p>(57) <b>ABSTRACT</b></p> <p>Compounds of the formula:</p> <div style="text-align: center;"></div> <p>wherein: n is an integer from 1-3; X is N, CH, provided that when X is N, n is 2 or 3; R is alkyl of 1 to 3 carbon atoms; R<sup>1</sup> is 2,4-diCl, 5-OMe; 2,4-diCl; 3,4,5-tri-OMe; 2-Cl, 5-OMe; 2-Me, 5-OMe; 2,4-di-Me; 2,4-diMe-5-OMe, 2,4-diCl, 5-OEt; R<sup>2</sup> is alkyl of 1 to 2 carbon atoms, and pharmaceutically acceptable salts thereof.</p>
---	--

12 Claims, No Drawings

Figure 1 Example of a pharmaceutical patent

In this example, the patent is on the therapeutic potential of the chemical structures bearing a 4-anilino-3-quinolinecarbonitrile core structure, depicted in Figure 2, with varying substituents at the phenyl ring and the 6 and 7 positions of the quinoline ring. The rules for substituents around the core structure are elaborated in prose below the core structure. These rules include, for example, the alkyl chain represented by (CH<sub>2</sub>)<sub>n</sub> can be of length of 1, 2 or 3; the group represented by X can be either N or CH, provided that when X is N the alkyl chain next to it is of length of 2 or 3 but not 1. The purpose of allowing patent applicant to use the Markush notation to make claims on similar structures is because of the well-known fact that similar structures often share similar properties and, in the medical field, similar pharmacological profiles. Given that the developer of a therapy often only has resources to progress one candidate out of a pool of similar structures for clinical testing,

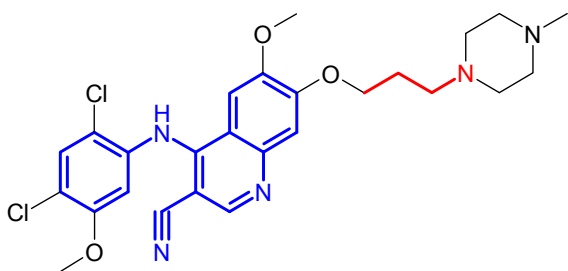
allowing the patent to cover the whole series of structures gives the applicant some protection against its competitors of attempting to develop a variant into a marketable product. On the other hand, patent examiners must balance the interests of patent applicants with those of the public. For example, if an application is so general that it covers any unrelated structures, its patentability must be challenged.



SMILES: N#CC1=C(NC2=CC=CC=C2)C2=CC=CC=C2N=C1

**Figure 2 Structure of the 4-anilino-3-quinolinecarbonitrile core**

Given a patent, a so-called substructure search can be used to check for a possible infringement[6]. Two examples with similar structures are shown in Figure 3. Both contain the 4-anilino-3-quinolinecarbonitrile substructure depicted in Figure 2 (highlighted in blue), but structure A is covered by the patent whereas B is not. In fact, A is bosutinib, the clinical candidate for treating breast cancer currently in Phase III studies[7]. The reason is because the patent claims that when X is N the neighboring alkyl chain must be 2 or 3 carbons in length (highlighted in red). Note that the 4-anilino-3-quinolinecarbonitrile substructure in the example is laid out differently from that in Figure 2. However, substructure searching algorithms should be able to detect the core structure by traversing the graph, which is a problem typically handled by a Chemistry software toolkit and is not in the scope of the proposed language.



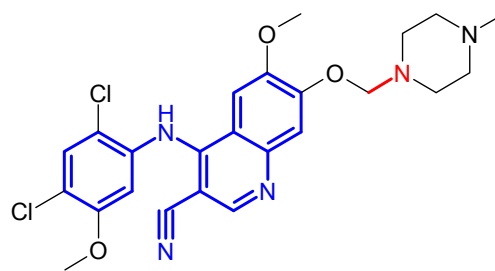
A

(Covered by US 7,417,148)

SMILES:

CN1CCN(CC1)CCCOC2=C(C=C3C(=C2)N=CC(=C3)NC4=CC(=C(C=C4Cl)Cl)OC)C#N)O

C



B

(Not covered by US 7,417,148)

SMILES:

COC1=CC(NC2=C(C=NC3=CC(OCN4CCN(C)CC4)=C(OC)C=C23)C#N)=C(Cl)C=C1Cl

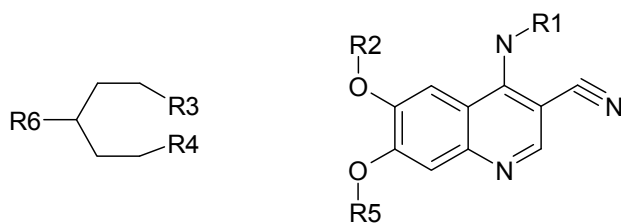
**Figure 3 Two examples demonstrating the coverage of patent US 7,417,148**

## II. Language Proposal

The current system used to describe claims in chemical inventions has many limitations. While the core in a Markush structure provides an unequivocal definition, the variations in the substituents have to be elaborated using an imprecise human language. Thus, when a possible infringement arises where a structure is suspected to be covered by an existing patent, legal professionals have to first determine the core of the patent is a substructure in the structure in question. If so, they need to then match each substituent of the structure against the variations claimed by the patent to determine if the entire structure is covered. Since variations can only be interpreted by human experts, this process is often tedious and error-prone. Similarly, when a patent application containing a Markush claim is reviewed, the patent examiner that is charged with determining the patentability often has a hard time manually searching the patent database and published literature for possible prior arts. For these reasons, the USPTO observed that searches of Markush claims “often consume a disproportionate amount of Office resources as compared to other types of claims[8].”

The Markush Description Language (MDL) proposed here is meant to make claims on variations of chemical structures machine-readable to enable automatic searches. Our language will borrow elements from the well-known Simplified Molecular Input Line Entry System (SMILES) [9] and its companion for chemical queries, SMARTS[10], of Daylight Chemical Information, Inc. SMILES and SMARTS are line notations for describing chemical structures as undirected graphs with weights denoting bond orders. Examples of SMILES for the 4-anilino-3-quinolinecarbonitrile core and for the two sample structures are displayed in Figure 2 and Figure 3 respectively. A web service exists that takes a SMILES string as input and generates a graphical depiction[11].

In this language, what we attempt to do is to separate the Markush structure into a fixed core component and a variable number of substituents. This is done by disconnecting the bonds linking a core and a substituent. Once a bond is disconnected, we cap the broken bond at the core end with a dummy R<sub>n</sub> atom and the corresponding substituent end with a dummy Z<sub>n</sub> atom where n is a unique integer assigned to the disconnected bond. The matching R and Z atoms allow us to reconnect a core to a substituent once it has been fully specified. The R and Z symbols are not part of SMILES/SMARTS and are reserved for this purpose only. For example, after the said bonds are disconnected, the core part of the Markush structure in US 7,417,148 will look like Figure 4.



SMILES: [R6]CCC([R3])CC[R4].[R1]NC1=C(C=NC2=CC(O[R5])=C(O[R2])C=C12)C#N

Figure 4 Core structured used in the Markush Description Language

## III. Language Syntax

MDL will support only two primitive types, *int* and *boolean*, and three complex types *String*, *Mol*, and, *QMol*. The *Mol* type will be used to describe real chemical structures such as A and B shown in Figure 3, and the *QMol* type will derive from *Mol* and will be used to describe query structures such as the one in Figure 4. The language will also support *tuples* natively, which contain immutable lists of any arbitrary objects.

MDL will provide an operator for concatenating two *QMol* objects together, which simply creates a single bond between the last atom of the structure on the left hand side and the first atom of the one on the right hand side. Furthermore, the word *covers* will be reserved as an operator which takes a *QMol* object and a *Mol* object and returns true if the structure on the right hand side is covered by the Markush claims defined by the query on the left hand side. Lastly, we will treat strings enclosed by triple quotes as chemical literals, namely SMILES or SMARTS.

To facilitate translation into Java in order to utilize a Java-based chemistry library for low-level operations, MDL will follow the imperative paradigm and will be statically typed.

In the code example below, the core object is the same as defined in Figure 4. The objects *r1*, *r2*, *x*, and *r* correspond to the respective substituents recited by the claims in US 7,417,148. Note that the group  $(\text{CH}_2)_n$  is represented by object *y*, which is joined with *x* as a whole to be connected with the core.

## IV. Code Example

```
QMol core =
  ''' [R6]CCC ([R3]) CC [R4] . [R1]NC1=C (C=NC2=CC (O [R5]) =C (O [R2]) C=C12) C#N''';

QMol [] r1 = (''' [Z1]c1c ([Cl]) cc ([Cl]) c (OC) c1''',
  ''' [Z1]c1c ([Cl]) cc ([Cl]) cc1''',
  ''' [Z1]c1cc (OC) c (OC) c (OC) c1''',
  ''' [Z1]c1c ([Cl]) ccc (OC) c1''',
  ''' [Z1]c1c (C) ccc (OC) c1''',
  ''' [Z1]c1c (C) cc (C) c (OC) c1''',
  ''' [Z1]c1c ([Cl]) cc ([Cl]) c (OCC) c1''');

QMol [] r2 = (''' [Z2]C''',
  ''' [Z2]CC''');

QMol [] x = (''' [Z3]C [Z4]''',
  ''' [Z3]N$ ([CR0] [CR0]) [Z4]''',
  ''' [Z3]N$ ([CR0] [CR0] [CR0]) [Z4]''');

QMol [] y = (''' C [Z5]''',
  ''' CC [Z5]''',
  ''' CCC [Z5]''');

// Now join the last atom of x with the first of y
QMol [] xy = X + Y;

QMol [] r = (''' [Z6]C''',
  ''' [Z6]CC''',
  ''' [Z6]CCC''',
  ''' [Z6]C (C) C''');

// Note the additions below are non-associative, namely the expression must
// be evaluated as (((core + r1) + r2) + xy) + r. Every time a core is
// joined with a substituent, one or more matching R-Z pairs are
// annihilated. After the last substituent, R, is joined, the Markush will
// have no Rs in it.
QMol [] markush = core + r1 + r2 + xy + r;

Mol structA =
  '''CN1CCN (CC1) CCCOC2=C (C=C3C (=C2) N=CC (=C3NC4=CC (=C (C=C4Cl) Cl) OC) C#N) OC''';
if (markush covers structA) {
  print "structA is covered by patent";
}

Mol structB =
  '''COc1=CC (NC2=C (C=NC3=CC (OCN4CCN (C) CC4) =C (OC) C=C23) C#N) =C (Cl) C=C1Cl''';
if (markush covers structB) {
  print "structB is covered by patent";
}
```

## Bibliography

1. Markush, E.A. 1924, Pharma-Chemical Corporation: USA.

2. *Ex parte Markush*. 340 O.G. 839. 1924.
3. Simmons, E.S., *Markush structure searching over the years*. World Patent Information, 2003. **25**: p. 195-202.
4. Cielen, E., *Searching Markush formulae directed to medical applications*. World Patent Information, 2009. **31**: p. 178-183.
5. Barnard, J.M. and P.M. Wright, *Towards in-house searching of Markush structures from patents*. World Patent Information, 2009. **31**: p. 97-103.
6. Leach, A.R. and V.J. Gillet, *An Introduction to Chemoinformatics*. 2005, Dordrecht, The Netherlands: Springer.
7. Vultur, A., et al., *SKI-606 (bosutinib), a novel Src kinase inhibitor, suppresses migration and invasion of human breast cancer cells*. Mol Cancer Ther, 2008. **7**(5): p. 1185-94.
8. *Examination of Patent Applications That Include Claims Containing Alternative Language*, P.a.T. Office, Editor. 2007, Federal Register. p. 44992-45000.
9. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*. Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.
10. *Daylight Theory Manual*, Daylight Chemical Information, Inc.: Aliso Viejo, CA.
11. *Interactive depiction of SMILES*. [cited 2011 February 8th]; Available from: <http://www.daylight.com/daycgi/depict>.