Supporting Information

ZINC – A Free Database of Commercially Available Compounds for Virtual Screening.

John J. Irwin and Brian K. Shoichet

Department of Pharmaceutical Chemistry

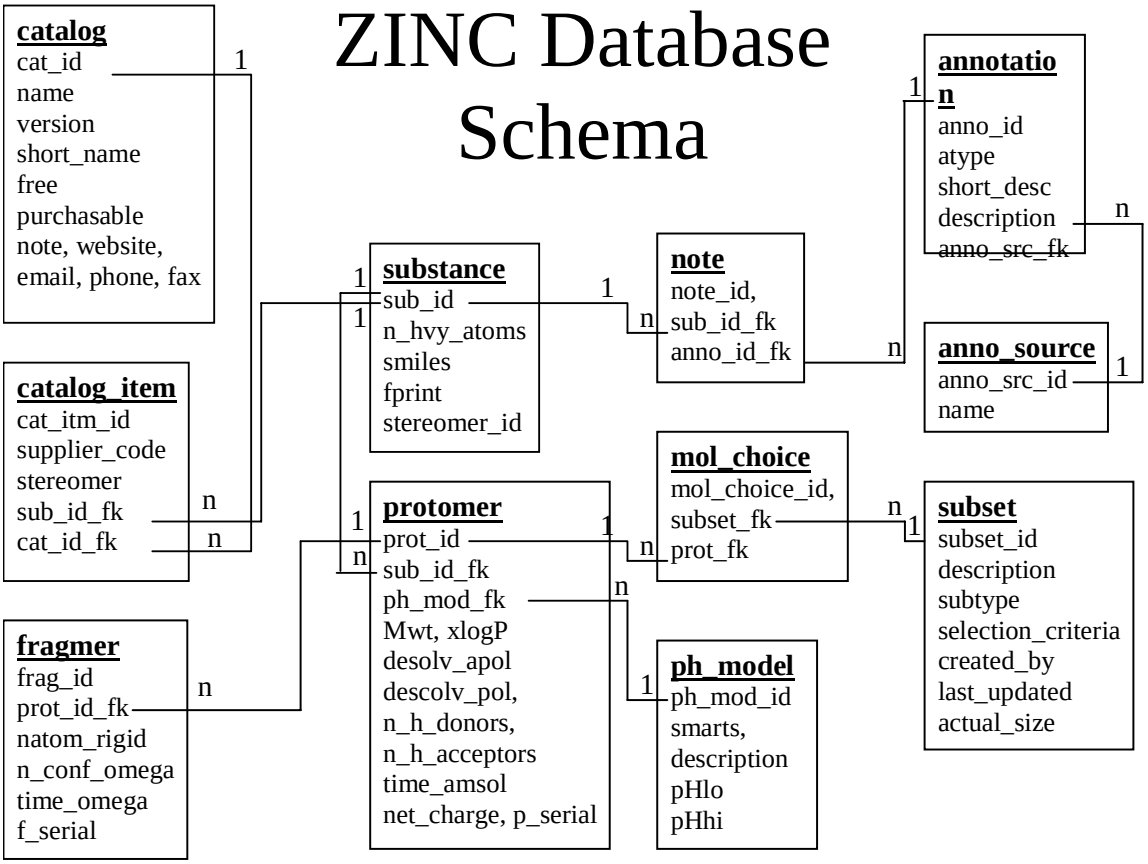
University of California San Francisco

Genentech Hall, 600 16th St, San Francisco CA 94143

Table S1. Current size of the MySQL tables used to hold the ZINC database.

Table	Purpose/example	# rows	Bytes	Total MB
substance	Unique chemical	877,941	400	330
protomer	Tautomer/protonation	986,965	400	380
fragmer	Rigid fragment choice	1,976,479	80	180
catalog	Source (e.g. Specs)	19	300	<1
catalog_item	Single buyable item	1,028,497	100	100
ph_model	e.g. pteridine, thiolate	7	200	<1

Figure S1. Schema of the ZINC database.



There are ten tables as follows.

1. Substance – Corresponds to a unique chemical species, i.e. a unique isomeric smiles string. Thus stereoisomers and configurational (E/Z) isomers have unique ZINC codes, even if they correspond to a single catalog entry (e.g. a racemic mixture). Molecules files that correspond to this level are available in isomeric SMILES format.

2. Catalog – Corresponds to a vendor-supplied file of molecules, usually dated and having a version number. Many vendors issue new catalogs monthly. We maintain two catalogs for each supplier: currently available compounds and retired compounds. We encourage our users to use the current catalog only.

3. Catalog_item – Corresponds to a single entry in a vendor's catalog, and has a single catalog number for purchasing. One catalog item may point to more than one substance, for example, a racemic mixture. A single substance may point to several catalog items, indicating its availability from those vendors.

4. Anno_source – Corresponds to sources of annotation, e.g. NCI, ChemBank and user-supplied annotations.

5. Annotation – Corresponds to a single annotation (e.g. COX-2 inhibitor).

6. Note – Is a single annotation that attributes a single annotation to a single substance.

7. Ph_model – Corresponds to our hand-crafted rules for alternate protonation forms of molecules (**table 5**).

8. Protomer – Corresponds to a protonated, 3D molecule, either a tautomer (e.g. folate) or an alternate protonated form (e.g. imidazole, pteridine). Molecule files that correspond to this level are available in mol2 and SDF formats from our website.

9. Fragmer – Corresponds to rigid fragment choices, also known as "anchors" in docking parlance. Typically rings are used for rigid fragments, so the number of fragmers is often the number of rigid ring systems. Molecule files that correspond to this level are available in DOCK flexi-base format.

10. Subset – Corresponds to selections of molecules from the database and available as exported files from our website. Subsets are also used for sets of molecules uploaded by users for processing.

Figure S2. Our version of the filter_light.txt parameter file for OpenEye's filter.1.0.2.

```
#This file defines the rules for filtering multi-structure files based on
#properties and substructure patterns.
#NOTE: this rule file is suitable for drug-like filtering and may be
#     too restrictive for reagent filtering.
#These rules include the contributions of countless chemists & modellers
including:
#Rule of 5:
#     Lipinski, C, et al., Adv.Drug Deliv. Rev., 23:3, 1997.
#General Filtering:
#     Oprea, T., JCAMD 14:251-264, 2000.
#XLogP:
#     Wang, R, Ying, Fu, & Lai, Luhua, J.Chem. Inf. Comput. Sci., 37:615-621,
1997.
#Polar Surface Area
#     Peter Ertl, Bernhard Rohde, & Paul Slzer, J. Med. Chem. 43:3714-3717,
2000.
#     David E. Clark, Journal of Pharmaceutical Sciences, 88(8):807-814,
1999.

MIN_MOLWT               60          "Minimum molecular weight"
MAX_MOLWT               600          "Maximum molecular weight"

MIN_SSSR_RINGS          0                "Minumum number of SSSR rings"
MAX_SSSR_RINGS          7                "Maximum number of SSSR rings"

MAX_RING_SIZE           12               "Maximum size of any SSSR ring"

MIN_CARBONS             3                "Minimum number of carbons"
MIN_HETEROATOMS         0                "Minimum number of heteroatoms"
MIN_Het_C_Ratio         0.00        "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio         2.0              "Maximum heteroatom to carbon
ratio"

#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) -
(BondsSharedWithOtherRings)
#must be >= 0, from JCAMD 14:251-265,2000.
ADJUST_ROT_FOR_RING_TRUE                 "BOOLEAN for weather to estimate
degrees of freedom in rings"
#ADJUST_ROT_FOR_RING_FALSE

MIN_ROT_BONDS           0                "Minimum number of rotatable bonds"
MAX_ROT_BONDS           12               "Maximum number of rotatable bonds"

MIN_RIGID_BONDS         0                "Minimum number of rigid bonds"
MAX_RIGID_BONDS         50               "Maximum number of rigid bonds"

MIN_HBOND_DONORS  0               "Minimum number of hydrogen-bond donors"

MAX_HBOND_DONORS  6               "Minimum number of hydrogen-bond donors"


MIN_HBOND_ACCEPTORS     0                "Minimum number of hydrogen-bond
acceptors"
```

```
MAX_HBOND_ACCEPTORS     10                "Minimum number of hydrogen-bond
acceptors"

MIN_COUNT_FORMAL_CRG        0       "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG          3                 "Maximum number of formal
charges"
MIN_SUM_FORMAL_CRG            -2                "Minimum sum of formal
charges"
MAX_SUM_FORMAL_CRG            2                 "Maximum sum of formal
charges"

MIN_XLOGP               -4.0              "Minimum XLogP"
MAX_XLOGP               6.0               "Maximum XLogP"

MIN_2D_PSA              0.0               "Minimum 2-Dimensional (SMILES)
Polar Surface Area"
MAX_2D_PSA              150.0       "Maximum 2-Dimensional (SMILES) Polar
Surface Area"

ALLOWED_ELEMENTS  H,B,C,N,O,F,S,P,Cl,Br,I
ELIMINATE_METALS Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#acceptable molecules must have <= instances of each of the patterns below

#specific, undesirable functional groups
RULE  0     Carbazides
RULE  0     Acid_anhydrides
RULE  0     Pentafluorophenyl_esters
RULE  0     Paranitrophenyl_esters
RULE  0     HOBT_esters
RULE  0     Triflates
RULE  0     Lawesson_s_reagent
RULE  0     Phosphoramides
RULE  0     Aromatic_azides
RULE  0     Beta_carbonyl_quart_nitrogen
#RULE 0     Acylhydrazide
RULE  0     Quarternary_C_Cl_I_P_or_S
RULE  0     Phosphoranes
RULE  0     Chloramidines
RULE  0     Nitroso
RULE  0     P_S_Halides
RULE  0     Carbodiimide
RULE  0     Isonitrile
RULE  0     Triacyloxime
RULE  0     Cyanohydrins
RULE  0     Acyl_cyanides
RULE  0     Sulfonyl_cyanides
RULE  0     Cyanophosphonates
RULE  0     Azocyanamides
RULE  0     Azoalkanals
RULE  0     Polyenes
RULE  0     Saponin_derivatives
RULE  0     Cytochalasin_derivatives
RULE  0     Cycloheximide_derivatives
RULE  0     Monensin_derivatives
RULE  0     Cyanidin_derivatives
RULE  0     Squalestatin_derivatives
```

```
#functional groups which often eliminate compounds from consideration
RULE  0     acid_halide
#RULE 0     aldehyde
#RULE 0     alkyl_halide
RULE  0     anhydride
RULE  0     azide
RULE  0     azo
RULE  0     di_peptide
RULE  0     long_aliphatic_chain               //(>7 atoms)
#RULE 0   michael_acceptor
RULE  0     beta_halo_carbonyl
#RULE 0   nitro
RULE  0     peroxide
RULE  0     phosphonic_acid
RULE  0     phosphonic_ester
RULE  0     phosphoric_acid
RULE  0     phosphoric_ester
RULE  0     sulfonic_acid
RULE  0     sulfonic_ester
RULE  0   triphenyl_phosphene
RULE  0     unbranched_chain                   //(>4_atoms)
#RULE 0     epoxide
#RULE 0     hetero_hetero
RULE  0     sulfonyl_halide
RULE  0     halopyrimidine
RULE  0     perhalo_ketone
#RULE 0     methyl_ketone
RULE  0     aziridine
#RULE 0     imine
RULE  0     oxalyl


#the dye group includes a set of patterns which describe all cpds with colors
in their names from the ACD98.2

RULE  0     dye

#functional groups which are allowed, but may not be wanted in high
quantities
#common functional groups

RULE  6     alcohol
RULE  8     alkene
RULE  4     amide
RULE  4     amino_acid
RULE  4     amine
RULE  4     primary_amine
RULE  4     secondary_amine
RULE  4     tertiary_amine
RULE  4     carboxylic_acid
RULE    6         halide
RULE  1     iodine
RULE  4     ketone
RULE  4     phenol

#other functional groups
```

```
RULE  4     alkyne
RULE  4     aniline
RULE  4     aryl_halide
RULE  4     carbamate
RULE  4     ester
RULE  4     ether
RULE  4     hydrazine
RULE  4     hydrazone
RULE  4     hydroxylamine
RULE  4     nitrile
RULE  4     sulfide
RULE  4     sulfone
RULE  4     sulfoxide
RULE  4     thiourea
RULE  4     thioamide
RULE  4     thiol
RULE  4     urea
```

Figure S3. Our version of the services/data/ionizer.ini file for Schrodinger's LigPrep.

```
##################################################################
# $Id: ionizer.ini,v 1.40 2004/02/06 19:46:18 reboul Exp $
# Control specs for ionization state expander
# Copyright 2002-2003 Schrodinger, LLC.
# All rights reserved.
##################################################################

##################################################################
#
# Ionization pattern spec syntax
# ==============================
#
# In the Ionizer's processing of this data file, blank lines
# are ignored, and any text to the right of a '#' is treated
# as a comment.
#
# All other text is treated as an ionization spec. Each valid
# ionization spec consists of a command name, followed by the
# command's required parameters. These ion spec commands are
# supported --
#
# Acid:     Specifies an acid group, and how to ionize it
#
# Base:     Specifies a Base group, and how to ionize it
#
# Exclude:  Specifies a group to exclude from consideration
#           as ionizable; used to selectively "forget" group
#           matches Based on prior Acid/Base specs
#
```

```
# Ionizable groups are specified using Schrodinger's linear
# substructure notation, which goes by the name "mmsubs". That
# syntax is described in this installed document --
#
#     $SCHRODINGER/services-v#####/doc/mmsubs_syntax.txt
#
################################################################
#
# Note Well
# =========
#
# Every input CT is tested for any substructures matching
# the specified ionizable groups, in the order in which they
# appear below. The effect of a later pattern match will
# supersede the effect of an earlier pattern match. Thus,
# anyone crafting pattern specs like these must think very
# carefully about the proper order for the desired effects.
#
# For example -- look below -- the sulfonic pattern has to
# come after the sulfinic pattern, because the sulfonic
# pattern is a specialization of the sulfinic pattern. Any
# group which matches sulfonic will also match sulfinic, but
# the opposite assertion is not necessarily true.
#
# If the order of the two patterns were reversed below, then
# a sulfonic match would always be superseded by the sulfinic
# match, thereby losing the intended distinction between the
# two.
#
# Because of the great generality of this syntax, the inter-
# pattern precedences can be much less obvious than for the
# sulfinic vs. sulfonic case. Be careful!
#
################################################################
#
# We now describe the command parameter syntax....
#
#   Acid  <mmsubs_pattern>  <ionized_fragment>  <pKa>
#   Base  <mmsubs_pattern>  <ionized_fragment>  <pKa>
#
# For each input CT, in any substructure found to match the
# specified mmsubs pattern, the atom matching the pattern's
# leading atom will be considered for ionization via fusion
# of the specified ionized fragment. The ion group will be
# treated as acid or Base according to the command used.
#
# The fragment named must be in this program's custom
# fragment library. The library is named "ionized", and it
# currently contains fragments named Ammonium, Hydroxide,
# N-minus, and Thiolate.
#
# The Acid/Base spec's 3rd parameter, a pKa value for the
# matching group, is used in deciding which ion combinations
# to actually generate in the output, Based on pK and/or
# pH considerations at run time. Note that pKa is used for
# both acids and Bases.
#
```

```
# You cannot specify an Acid/Base pattern in this file
# without supplying a pKa value for it. It is assumed that
# a pKa value is at least somewhat accurate, as it is going
# to be used in the program's restriction determinations.
#
# When one Acid/Base pattern is a specialization of another
# Acid/Base pattern, it is probably the case that the pKa
# value for the more specialized pattern can be given more
# precisely. Assuming that the patterns are presented in
# order of increasing specificity, the more precise pKa will
# be assigned ultimately.
#
# Exclude specs don't specify a fragment or pKa....
#
#   Exclude  <mmsubs_pattern>
#
# For each input CT, in any substructure matching the
# specified pattern, the atom matching the pattern's
# leading atom will be _excluded_ from consideration for
# ionization.
#
# An Exclude pattern match causes an effect only if the
# leading atom coincides with a prior Acid/Base pattern
# match's leading atom. If there is no corresponding prior
# Acid/Base match, the Exclude match is simply ignored.
#
# Because of the Ionizer's top-to-bottom processing of these
# pattern specs, each Exclude pattern must be placed after
# all the Acid/Base specs whose matches it might negate.
#
# It is possible for a given input CT atom to be matched as
# an ionization center due to some Acid/Base pattern spec,
# then excluded from consideration due to a later Exclude
# pattern, then re-matched due to an even later Acid/Base
# spec.
#
################################################################
#
# Important note
# ==============
#
# The following specifications are _not_ an encyclopedic list
# of ionizable groups!
#
# Some users will wish to prepare their own customized data
# file, presumably by adapting a copy from this one, and then
# running with the customized data, specified via command-
# line option.
#
# Such users must understand Schrodinger's "mmsubs" linear
# substructure notation. Correct use of the syntax is not
# trivially easy. Users may need to contact Schrodinger for
# assistance.
#
################################################################

################################################################
```

```
# Specs for acids, to be deprotonated

# carboxylic
#Acid O0(-H0)-C2(=O0)-C0 Hydroxide 4.0
# for now, I don"t want them to be protonated for the CCP pocket
Acid O0(-H0)-C2(=O0)-C0 Hydroxide 2.0

# phosphoric
Acid O0(-H0)-P0(=O0)(-O0)-O0 Hydroxide 2.1

# phosphorylamide
Acid O0(-H0)-P0(=O0)(-N0)-O0 Hydroxide 2.6

# phosphonic
Acid O0(-H0)-P0(=O0)(-O0)-C0 Hydroxide 2.5

# phosphonamide
Acid O0(-H0)-P0(=O0)(-N0)-C0 Hydroxide 2.9

# sulfuric
Acid O0(-H0)-S0(=O0)(=O0)-O0 Hydroxide -2.0

# sulfinic
Acid O0(-H0)-S0(=O0)-C0 Hydroxide 2.0

# sulfonic
# (supersedes match on sulfinic above)
Acid O0(-H0)-S0(=O0)(=O0)-C0 Hydroxide -1.0

# hydroxamic
Acid O0(-H0)-N0(-H0)-C0=O0 Hydroxide 8.5
Acid O0(-H0)-N0(-C3)-C0=O0 Hydroxide 8.5 # estimated

# sulfonamides
#   aromatic or alkene
Acid N0(-H0)(-C2=C2)-S0(=O0)(=O0)-C0 N-minus 8.2
#   pyridyl
Acid N0(-H0)(-C2=N2)-S0(=O0)(=O0)-C0 N-minus 8.2
#   carbonyl
Acid N0(-H0)(-C2=O2)-S0(=O0)(=O0)-C0 N-minus 4.5
#   alkyl
Acid N0(-H0)(-C3)-S0(=O0)(=O0)-C0 N-minus 11.6
#   Chlorine
Acid N0(-H0)(-Cl)-S0(=O0)(=O0)-C0 N-minus 4.5

# tetrazole
Acid N0(-H0)-N0=N0-N0=C2-1 N-minus 4.5
Acid N0(-H0)-N0=N0-C2=N0-1 N-minus 4.5

# phenol
Acid O3(-H0)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*3 Hydroxide 10.0

# Regarding the phenol pattern, one co-worker said --
#
# "The phenol pattern above correctly works with most of my tested
# phenols and non-phenols. It is great for single carbon aromatic
# rings, and avoids greedy matches with compounds it should not.
```

```
# However, the pattern above does not match some canonical resonance
# structures of polycyclic benzenoid aromatics."
#
# "My opinion is that the pattern is pretty good, and certainly
# better than not having it at all. The polycyclic benzenoid
# compounds are vexing."

# 2-nitrosophenol/2-nitrophenol
Acid O0(-H0)-C2*C2(-N0=O0)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*3 Hydroxide 6.5

# 4-nitrosophenol/4-nitrophenol
Acid O0(-H0)-C2*C2(-00)*C2(-00)*C2(-N0=O0)*C2(-00)*C2(-00)*3 Hydroxide 6.5

# 3,5-dinitrosophenol/3,5-dinitrophenol
Acid O0(-H0)-C2*C2(-00)*C2(-N0=O0)*C2(-00)*C2(-N0=O0)*C2(-00)*3 Hydroxide 6.7

# alkylthiol
Acid S0(-H0)-C3 Thiolate 9.5

# thiophenol
Acid S0(-H0)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*3 Thiolate 6.6
Acid S0(-H0)-C2*N2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*3 Thiolate 6.6
Acid S0(-H0)-C2*C2(-00)*N2*C2(-00)*C2(-00)*C2(-00)*3 Thiolate 6.6
Acid S0(-H0)-C2*C2(-00)*C2(-00)*N2*C2(-00)*C2(-00)*3 Thiolate 6.6
Acid S0(-H0)-C2*N2*C2(-00)*N2*C2(-00)*C2(-00)*3 Thiolate 6.6

# Don't match sulfinic acid tautomer
Exclude S0(-H0)(=O0)(=O0)

################################################################
# In many ring patterns, both above and below here, note the
# use of *C2(-00)*, where -00 (two zeroes) signifies single
# bond to any atom. This restricts the matches to only those
# substructures with aromatic carbons. We have to do this in
# light of our atom types. As one co-worker explained --
#
# "[B]ecause of our broad definition of the aromatic C (*C2*),
# our patterns would [otherwise] match quinone or uracyl type
# compounds. Hence we [...] have to impose a single-bonded
# substituent on every aromatic C to distinguish between them
# and carbonyl type C2."
################################################################

################################################################
# Specs for Bases, to be protonated

# We require Base pattern leading Nitrogens to be uncharged,
# to ignore some input molecules' N+ atoms otherwise matching
# these patterns; hence the "[Ajay #79]" qualifiers, which are the
# mmsubs-extension syntax for zero formal charge.


# dialkylaniline
Base N0{0}(-C3)(-C3)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*4 Ammonium
4.5

# amine
```

```
Base N0{0}(-H0)(-H0)-C3 Ammonium 10.5 # primary
Base N0{0}(-H0)(-C3)-C3 Ammonium 11.0 # secondary
Base N0{0}(-C3)(-C3)-C3 Ammonium 10.0 # tertiary

# hydrazine
Base N0{0}(-H0)(-H0)-N3 Ammonium 10.0 # primary
Base N0{0}(-H0)(-C3)-N3 Ammonium 10.0 # secondary
Base N0{0}(-C3)(-C3)-N3 Ammonium 10.0 # tertiary

# imine
Base N0{0}(-H0)=C2(-H0)-C3 Ammonium 11.5
Base N0{0}(-H0)=C2(-C3)-C3 Ammonium 11.5
Base N0{0}(-C3)=C2(-H0)-C3 Ammonium 11.5
Base N0{0}(-C3)=C2(-C3)-C3 Ammonium 11.5

# amidine
Base N0{0}(-H0)=C2(-H0)-N0-H0 Ammonium 12.0
Base N0{0}(-H0)=C2(-H0)-N0-C3 Ammonium 12.0
Base N0{0}(-H0)=C2(-C0)-N0-H0 Ammonium 12.0
Base N0{0}(-H0)=C2(-C0)-N0-C3 Ammonium 12.0
Base N0{0}(-C3)=C2(-H0)-N0-H0 Ammonium 12.0
Base N0{0}(-C3)=C2(-H0)-N0-C3 Ammonium 12.0
Base N0{0}(-C3)=C2(-C0)-N0-H0 Ammonium 12.0
Base N0{0}(-C3)=C2(-C0)-N0-C3 Ammonium 12.0

# guanidine
Base N0{0}(-H0)=C2(-N0-H0)-N0-H0 Ammonium 12.5
Base N0{0}(-H0)=C2(-N0-H0)-N0-C3 Ammonium 12.5
Base N0{0}(-H0)=C2(-N0-C3)-N0-H0 Ammonium 12.5
Base N0{0}(-H0)=C2(-N0-C3)-N0-C3 Ammonium 12.5
Base N0{0}(-C3)=C2(-N0-H0)-N0-H0 Ammonium 12.5
Base N0{0}(-C3)=C2(-N0-H0)-N0-C3 Ammonium 12.5
Base N0{0}(-C3)=C2(-N0-C3)-N0-H0 Ammonium 12.5
Base N0{0}(-C3)=C2(-N0-C3)-N0-C3 Ammonium 12.5

# enamine
Base N0{0}(-H0)(-H0)-C2(-H0)=C2 Ammonium 10.5 # primary
Base N0{0}(-H0)(-H0)-C2(-C3)=C2 Ammonium 10.5 # primary
Base N0{0}(-H0)(-C3)-C2(-H0)=C2 Ammonium 10.5 # primary
Base N0{0}(-H0)(-C3)-C2(-C3)=C2 Ammonium 10.5 # primary
Base N0{0}(-C3)(-C3)-C2(-H0)=C2 Ammonium 10.5 # primary
Base N0{0}(-C3)(-C3)-C2(-C3)=C2 Ammonium 10.5 # primary
#
# Screen out some matches on enamines above --
#
Exclude     N0{0}(-H0)(-H0)-C2=C2-C2=O0
Exclude     N0{0}(-H0)(-C3)-C2=C2-C2=O0
Exclude     N0{0}(-C3)(-C3)-C2=C2-C2=O0
#
Exclude     N0{0}(-H0)(-H0)-C2=C2-C1%N0
Exclude     N0{0}(-H0)(-C3)-C2=C2-C1%N0
Exclude     N0{0}(-C3)(-C3)-C2=C2-C1%N0


# aniline
Base N0{0}(-H0)(-H0)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*4 Ammonium
4.5 # primary
```

```
Base N0{0}(-H0)(-C0)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*4 Ammonium
4.5 # secondary
#the problem with the rule above is that compounds also get protonated, if
the N is part of a ring
Exclude N0*C2*C2*C2*C2*1            #anniliertes pyrrole
Exclude N0*C2*N0*C2*C2*1            #anniliertes imidazole
Exclude N0*N0*C2*C2*C2*1            #anniliertes pyrazole (?)
#Exclude F or Cl sub.
Exclude N0(-H0)(-00)-C2*C2(-Hl)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*4    # F or
Cl substiuent
Exclude N0(-H0)(-00)-C2*C2(-00)*C2(-Hl)*C2(-00)*C2(-00)*C2(-00)*4    # F or
Cl substiuent
Exclude N0(-H0)(-00)-C2*C2(-00)*C2(-00)*C2(-Hl)*C2(-00)*C2(-00)*4    # F or
Cl substiuent
Exclude N0(-H0)(-00)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-Hl)*C2(-00)*4    # F or
Cl substiuent
Exclude N0(-H0)(-00)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-Hl)*4    # F or
Cl substiuent


Base N0{0}(-C0)(-C0)-C2*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*4 Ammonium
4.5 # tertiary

# 1,4-diaminobenzene
Base N0{0}(-H0)(-H0)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*4
Ammonium 6.2
Base N0{0}(-H0)(-H0)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.2
Base N0{0}(-H0)(-H0)-C2*C2(-00)*C2(-00)*C2(-N0(-C3)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.2
Base N0{0}(-H0)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*4
Ammonium 6.0
Base N0{0}(-H0)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.0
Base N0{0}(-H0)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-C3)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.0
Base N0{0}(-C3)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*4
Ammonium 6.0
Base N0{0}(-C3)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-H0)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.0
Base N0{0}(-C3)(-C3)-C2*C2(-00)*C2(-00)*C2(-N0(-C3)(-C3))*C2(-00)*C2(-00)*4
Ammonium 6.0

# N-heterocycles....

# imidazole
Base N0{0}=C2-N0-C2=C2-1 Ammonium 7.0
#RB no Cl or F substiuents
Exclude N0=C2(-Hl)-N0-C2=C2-1
Exclude N0=C2-N0-C2=C2(-Hl)-1
Exclude N0=C2-N0-C2(-Hl)=C2-1

# pyridine
Base N2{0}*C2(-00)*C2(-00)*C2(-00)*C2(-00)*C2(-00)*1 Ammonium 5.5

#3-aminopyridine has got an pKa of 6.2 in ACD labs
#RB
```

```
Base N2{0}*C2(-00)*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*C2(-00)*1 Ammonium 6.2

# pyridine with C=O, C=C or halide
Exclude N2*C2(-Hl)*C2*C2*C2*C2*1
Exclude N2*C2*C2(-Hl)*C2*C2*C2*1
Exclude N2*C2*C2*C2(-Hl)*C2*C2*1
Exclude N2*C2(*C2=O2)*C2*C2*C2*C2*1
Exclude N2*C2*C2(*C2=O2)*C2*C2*C2*1
Exclude N2*C2*C2*C2(*C2=O2)*C2*C2*1
Exclude N2*C2(*C2=N2)*C2*C2*C2*C2*1
Exclude N2*C2*C2(*C2=N2)*C2*C2*C2*1
Exclude N2*C2*C2*C2(*C2=N2)*C2*C2*1

# 2-aminopyridine (ACD labs says 6.7) (they should still be protonated at pH
4.5, even with F or Cl subs.)
Base N2{0}*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*C2(-00)*C2(-00)*1 Ammonium 7.5
Base N2{0}*C2(-N0(-H0)(-C3))*C2(-00)*C2(-00)*C2(-00)*C2(-00)*1 Ammonium 7.5
Base N2{0}*C2(-N0(-C3)(-C3))*C2(-00)*C2(-00)*C2(-00)*C2(-00)*1 Ammonium 7.5


# 4-aminopyridine (they should still be protonated at pH 4.5, even with F or
Cl subs. the better way to go would probably to define diffrent pkas
#for the subtituet ones)
Base N2{0}*C2(-00)*C2(-00)*C2(-N0(-H0)(-H0))*C2(-00)*C2(-00)*1 Ammonium 9.0
Base N2{0}*C2(-00)*C2(-00)*C2(-N0(-H0)(-C3))*C2(-00)*C2(-00)*1 Ammonium 9.0
Base N2{0}*C2(-00)*C2(-00)*C2(-N0(-C3)(-C3))*C2(-00)*C2(-00)*1 Ammonium 9.0

# 4-methoxypyridine
Base N2{0}*C2(-00)*C2(-00)*C2(-O0-C3)*C2(-00)*C2(-00)*1 Ammonium 6.5

# new rules for 3-aminopyridazine (pyridazine itself has got a pKa of about
2.5)
Base N2{0}*N2*C2(-N0(-H0)(-H0))*C2*C2*C2*1 Ammonium 5.0
#No substituentns on the N-N pattern
Exclude N2(*00)*N2*C2(-N0(-H0)(-H0))*C2*C2*C2*1
Exclude N2*N2(*00)*C2(-N0(-H0)(-H0))*C2*C2*C2*1
#Exclude 3,5, Diaminopyridazines, since they would be again covered by the
following rule (RB)
Exclude N2*N2*C2(-N0(-H0)(-H0))*C2*C2(-N0(-H0)(-H0))*C2*1

# new rules for 4-aminopyridazine (pyridazine itself has got a pKa of about
2.5)
Base N2{0}*N2*C2*C2(-N0(-H0)(-H0))*C2*C2*1 Ammonium 6.5
Exclude N2(*00)*N2*C2*C2(-N0(-H0)(-H0))*C2*C2*1
Exclude N2*N2(*00)*C2*C2(-N0(-H0)(-H0))*C2*C2*1

# new rules for 2-aminopyrazine (pyrazine itself has got a pKa of about -0.5)
Base N2{0}*C2(-N0(-H0)(-H0))*C2*N2*C2*C2*1 Ammonium 3.0

#new rule for 4-aminopyrimidne
#Base N2{0}*C2(-00)*N0*C2(-N0(-H0)(-H0))*C2*C2*1 Ammonium 5.5


# new rule for 2,4-aminopyrimidine (2-aminopyrimidine has got a pKa of about
2.6)
Base N2{0}*C2(-N0(-H0)(-H0))*N2*C2(-N0(-H0)(-H0))*C2*C2*1 Ammonium 6.5
```

```
#exclude pyrimidines with a Cl or F substituent (RB)
Exclude N2*C2(-Hl)*N0*C2(-N0(-H0)(-H0))*C2*C2*1
Exclude N2*C2*N0*C2(-N0(-H0)(-H0))*C2(-Hl)*C2*1
Exclude N2*C2*N0*C2(-N0(-H0)(-H0))*C2*C2(-Hl)*1

#exclude all 6-membered N-containing rings with more than 1 F or Cl (RB)
Exclude N0*00(-Hl)*00(-Hl)*00*00*00*1
Exclude N0*00(-Hl)*00*00(-Hl)*00*00*1
Exclude N0*00(-Hl)*00*00*00(-Hl)*00*1
Exclude N0*00(-Hl)*00*00*00*00(-Hl)*1
Exclude N0*00*00(-Hl)*00(-Hl)*00*00*1
Exclude N0*00*00(-Hl)*00*00(-Hl)*00*1



#new rule for 2amino-thiazole
Base N2{0}*C2(-N0(-H0)(-H0))*S0*C2(-00)*C2(-00)*1 Ammonium 5.5


# purine
# (supersedes match on imidazole above) and other heterocycles (new
placement, RB)
#Exclude N0=C2-N0-C2*N0*C2(-00)*N0*C2*C2(-1)(*5)
#Exclude N2*C2*N2*C2*C2*1
Exclude N0*C0*N0*C0*N0*C0*N0*C0*C0(*1)(*4)
#Why does this rule not get applied?????


# benzimidazole family
Exclude N2{0}*C2*N2*C2*C2(*02*02*02*02*4)*1

################################################################
# Alternative method for not matching N+ atoms
#
# Don't specify "{0}" in the Base N* patterns above, but
# then, after all the Base N* patterns, Exclude matches on
# Nitrogens with total bond order 4, which must be N+ atoms.
#
# With "{0}" qualifiers still in place above, the following
# is redundant, that is, it has no effect.
#
# Because the following patterns are so simple, they may hit
# some Nitrogens not matched by the Base patterns above. Any
# such matches are simply ignored, since there are no prior
# matches to undo.

Exclude N0(-00)(-00)(-00)(-00)    # 4 single bonds
Exclude N0(-00)(-00)(=00)         # 2 single + 1 double
Exclude N0(=00)(=00)              # 2 double
Exclude N0(-00)(%00)              # 1 single + 1 triple

# exclude amides
Exclude  N0*C0-O0-H0           # one aromatic, one OH
Exclude  N0-C0=O0           # one aromatic, one OH

################################################################
# Please do not alter or remove the comments below capturing
```

```
# the RCS revision info on this data file! The Ionizer's
# handling of the -v|-ver|-version option expects to find
# these here, in exactly this form --
#
# VERSION $Revision: 1.40 $
# VERSION $Date: 2004/02/06 19:46:18 $
###############################################################
```

Figure S4. Our version of the macromodel/data/tautomer_list file for Schrodinger's LigPrep.

```
Note that we have changed the ligprep script to use 20% (0.2) as the lower
limit for tautomeric equilibria. Schrodinger's default value was 0.01.

# instructions:
# Within a set of tautomers
# Covalent bonds involving Hydrogen that are moved in any tautomer must be
represented
# explicitly and proceeded by a -.
# Hydrogens that move must be respresented using a [ #1]
# Bonds that change in any tautomer must be represented explicitly.
# Heavy atoms must appear in the same order in all tautomers.
# All instances of asymmetric tautomers must be represented
# (see for example double-sided_ketol-enol).

# larger matches have precedence
# earlier matches of the same size have precedence

# for changing in bonding for aromatic N's (n)
# all bond orders which should change need to be represented by ':'
# see imidazole for an example

tautomer_set{
    name: imidazole

    tautomer{
      name: form1
      pattern: c1n(-[ #1])-c=[nX2]c1
      probability: 0.50
    }

    tautomer{
      name: form2
      pattern: c1[nX2]=c-n(-[ #1])c1
      probability: 0.50
    }
}

tautomer_set{
    name: double-ket-enol
# may want to make special allowances for outside enol (C=C outside
```

```
# central region)
# From: Handbook of organic chemistry

    tautomer{
      name: enol_ket
      pattern: [CX3](-[*;!#8;!#7])(=[CX3]([*;!#8;!#7])-O-[ #1])-
[CX3]([*;!#8;!#7])=O
      probability: 0.381
    }

    tautomer{
      name: diket
      pattern: C(-[ #1])(-[*;!#8;!#7])(-[CX3]([*;!#8;!#7])=O)-
[CX3]([*;!#8;!#7])=O
      probability: 0.238
    }

    tautomer{
      name: ket_enol
      pattern: [CX3](-[*;!#8;!#7])(-[CX3]([*;!#8;!#7])=O)=[CX3]([*;!#8;!#7])-
O-[ #1]
      probability: 0.381
    }
}

tautomer_set{
    name: single-sided_ket-enol
# From: Handbook of organic chemistry

    tautomer{
      name: enol
      pattern: [CX3]=[CX3]([*;!#8;!#7])-O-[ #1]
      probability: 0.00005
    }

    tautomer{
      name: ket
      pattern: [CX4](-[ #1])-[CX3]([*;!#8;!#7])=O
      probability: 0.99995
    }
}

tautomer_set{
    name: double-sided_ket-enol
# From: Handbook of organic chemistry

    tautomer{
      name: 1enol
      pattern: [CX3]=[CX3](-[CX4]-[ #1])-O-[ #1]
      probability: 0.00000001
    }

    tautomer{
      name: ket
      pattern: [CX4](-[ #1])-[CX3](-[CX4]-[ #1])=O
      probability: .99999998
    }
```

```
    tautomer{
      name: 2enol
      pattern: [CX4](-[ #1])-[CX3](=[CX3])-O-[ #1]
      probability: 0.00000001
    }
}

tautomer_set{
    name: single-sided_thio_ket-enol

    tautomer{
      name: thiol
      pattern: [CX3]=[CX3]([*;!#8;!#7])-S-[ #1]
      probability: 0.00005
    }

    tautomer{
      name: thio-acetal
      pattern: [CX4](-[ #1])-[CX3]([*;!#8;!#7])=S
      probability: 0.99995
    }
}

tautomer_set{
    name: double-sided_thio_ket-enol

    tautomer{
      name: 1thiol
      pattern: [CX3]=[CX3](-[CX4]-[ #1])-S-[ #1]
      probability: 0.00005
    }

    tautomer{
      name: thio-ketone
      pattern: [CX4](-[ #1])-[CX3](-[CX4]-[ #1])=S
      probability: 0.9999
    }

    tautomer{
      name: 2thiol
      pattern: [CX4](-[ #1])-[CX3](=[CX3])-S-[ #1]
      probability: 0.00005
    }
}

tautomer_set{
    name: single-sided_N

    tautomer{
      name: enol
      pattern: [CX3]=[CX3]([*;!#8;!#7])-[NX3](-[*])-[ #1]
      probability: 0.999986
    }

    tautomer{
      name: ket
```

```
      pattern: [CX4](-[ #1])-[CX3]([*;!#8;!#7])=[NX2](-[*])
      probability: 0.000014
    }
}

tautomer_set{
    name: double-sided_N

    tautomer{
      name: 1enol
      pattern: [CX3]=[CX3](-[CX4]-[ #1])-[NX3](-[*])-[ #1]
      probability: 0.499996
    }

    tautomer{
      name: ket
      pattern: [CX4](-[ #1])-[CX3](-[CX4]-[ #1])=[NX2]-[*]
      probability: 0.000008
    }

    tautomer{
      name: 2enol
      pattern: [CX4](-[ #1])-[CX3](=[CX3])-[NX3](-[*])-[ #1]
      probability: 0.499996
    }
}

tautomer_set{
    name: 1,2-diazole

    tautomer{
      name: 1H
      pattern: c-1=c-c=[nX2]-[nX3](-[ #1])-1
      probability: 0.50
    }
    tautomer{
      name: 2H
      pattern: c=1-c=c-[nX3](-[ #1])-[nX2]=1
      probability: 0.50
    }
}


tautomer_set{
    name: triazole

    tautomer{
      name: 1H
      pattern: c-1=c-[nX2]=[nX2]-[nX3](-[ #1])-1
      probability: 0.25
    }
    tautomer{
      name: 2H
      pattern: c=1-c=[nX2]-[nX3](-[ #1])-[nX2]=1
      probability: 0.5
    }
    tautomer{
```

```
      name: 3H
      pattern: c-1=c-[nX3](-[ #1])-[nX2]=[nX2]-1
      probability: 0.25
    }
}

tautomer_set{
    name: tetraazole

    tautomer{
      name: 1H
      pattern: c=1-[nX3](-[ #1])-[nX2]=[nX2]-[nX2]=1
      probability: 0.40
    }

    tautomer{
      name: 2H
      pattern: c-1=[nX2]-[nX3](-[ #1])-[nX2]=[nX2]-1
      probability: 0.10
    }

    tautomer{
      name: 3H
      pattern: c=1-[nX2]=[nX2]-[nX3](-[ #1])-[nX2]=1
      probability: 0.10
    }

    tautomer{
      name: 4H
      pattern: c-1=[nX2]-[nX2]=[nX2]-[nX3](-[ #1])-1
      probability: 0.40
    }
}


tautomer_set{
    name: 4-oxy-pyridine
    tautomer{
      name: one
      pattern: [nX3]-1(-[ #1])-c=c-c(=O)-c=c-1
      probability: 0.50
    }
    tautomer{
      name: hydroxy
      pattern: [nX2]=1-c=c-c(-O-[ #1])=c-c=1
      probability: 0.50
    }
}
tautomer_set{
    name: 2-oxy-pyridine
# see if non-one or hydroxy c can be changed to a
    tautomer{
      name: one
      pattern: [nX3]-1(-[ #1])-c(=O)-[a]=[a]-[a]=[a]-1
#      pattern: [nX3]-1(-[ #1])-c(=O)-c=c-c=c-1
      probability: 0.90
    }
```

```
    tautomer{
      name: hydroxy
      pattern: [nX2]=1-c(-O-[ #1])=[a]-[a]=[a]-[a]=1
#      pattern: [nX2]=1-c(-O-[ #1])=c-c=c-c=1
      probability: 0.10
    }
}


# disabled because not all forms interconvert readily
#tautomer_set{
#    name: buten_aldehyde
#
#    tautomer{
#      name: but-3-en-aldehyde
#      pattern: [CX3](=O)-[CX4](-[ #1])-[CX3]=[CX3](!#8)!#8
#      probability: 0.005
#    }
#    tautomer{
#      name: but-2-en-aldehyde
#      pattern: [CX3](=O)-[CX3]=[CX3]-[CX4](-[ #1])(!#8)!#8
#      probability: 0.99
#    }
#    tautomer{
#      name: but-1,3-dien-1-ol
#      pattern: [CX3](-O-[ #1])=[CX3]-[CX3]=[CX3](!#8)!#8
#      probability: 0.005
#    }
#}

# disabled because not all forms interconvert readily
#tautomer_set{
#    name: penten_aldehyde
#
#    tautomer{
#      name: pent-3-en-aldehyde
#      pattern: [CX3](=O)-[CX4](-[ #1])-[CX3]=[CX3]-[CX3]=O
#      probability: 0.45
#    }
#    tautomer{
#      name: pent-3-en-1-aldehyde-5-ol
#      pattern: [CX3](-O-[ #1])=[CX3]-[CX3]=[CX3]-[CX3]=O
#      probability: 0.05
#    }
#    tautomer{
#      name: pent-3-en-aldehyde
#      pattern: [CX3](=O)-[CX3]=[CX3]-[CX4](-[ #1])-[CX3]=O
#      probability: 0.45
#    }
#    tautomer{
#      name: pent-3-en-aldehyde
#      pattern: [CX3](=O)-[CX3]=[CX3]-[CX3]=[CX3]-O-[ #1]
#      probability: 0.05
#    }
#}

tautomer_set{
```

```
    name: guanine
# Tuatomer naming from Hanus et a. JACS 125, 7678 (2003)
# Their frame work atom numbering:
# (bond orders left out)
#
#              O
#              |
#        N     C
#      /   \ /   \
#     C     C     N
#     |     |     |
#     N-----C     C
#             \   / \
#               N    N
#
#
#              10
#              |
#        7     6
#      /   \ /   \
#     8     5     1
#     |     |     |
#     9-----4     2
#             \   / \
#               3    N
# So that tautomer names equivalences are:
# Used Here      Hanus et al.
#  1-5           1-9
#  1-7           1-7
#  3-7           3-7
#  3-5           3-9
#  5-7           7-9
#  5-O           9-O
#  7-O           7-O
    tautomer{
      name: 1H,9H,2-amino,6-one
      pattern: [nX3]-1(-[ #1])-c(-[NX3]-[ #1])=[nX2]-c=2[nX3](-[ #1])-
c=[nX2]c=2-c-1=O
      probability: 0.7781
    }
    tautomer{
      name: 1H,7H,2-amino,6-one
      pattern: [nX3]-1(-[ #1])-c(-[NX3]-[ #1])=[nX2]-c=2[nX2]=c-[nX3](-[
#1])c=2-c-1=O
      probability: 0.2194
    }
    tautomer{
      name: 3H,7H,2-amino,6-one
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX3](-[ #1])-c=2[nX2]=c-[nX3](-[
#1])c=2-c-1=O
      probability: 0.0025
    }
    tautomer{
      name: 3H,9H,2-amino,6-one
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX3](-[ #1])-c=2[nX3](-[ #1])-
c=[nX2]c=2-c-1=O
      probability: 1.67e-05
```

```
    }
    tautomer{
      name: 9H,2-amino,6-hydroxy
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX2]=c-2[nX3](-[ #1])-c=[nX2]c-2=c-1-
O-[ #1]
      probability: 2.39e-06
    }
    tautomer{
      name: 7H,2-amino,6-hydroxy
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX2]=c-2[nX2]=c-[nX3](-[ #1])c-2=c-1-
O-[ #1]
      probability: 4.11e-7
    }
    tautomer{
      name: 1H,3H,7H,2-imino,6-one
      pattern: [nX3]-1(-[ #1])-c(=[NX2])-[nX3](-[ #1])-c=2[nX2]=c-[nX3](-[
#1])c=2-c-1=O
      probability: 2.34e-7
    }
    tautomer{
      name: 1H,3H,9H,2-imino,6-one
      pattern: [nX3]-1(-[ #1])-c(=[NX2])-[nX3](-[ #1])-c=2[nX3](-[ #1])-
c=[nX2]c=2-c-1=O
      probability: 2.18e-8
    }
#    tautomer{
#      name: 1H,7H,2-imino,6-hydroxy
#      pattern: [nX3]-1(-[ #1])-c(=[NX2])-[nX2]=c-2[nX2]=c-[nX3](-[ #1])c-
2=c-1-O-[ #1]
#      probability:  2.63e-19
#    }

}

tautomer_set{
    name: guanosine
# Actually a methyl is used in place of the sugar which is OK for
# tautomers.
# This is similar to guanine except that there is one less H atom
# to move around and position 9 is blocked by the methyl
#
#
#              10
#              |
#        7     6
#      /   \ /   \
#     8     5     1
#     |     |     |
#     9-----4     2
#    /       \   / \
#   Me         3    N

# based upon guanine

    tautomer{
      name: 1H,2-amino,6-one
```

```
      pattern: [nX3]-1(-[ #1])-c(-[NX3]-[ #1])=[nX2]-c=2[nX3](-[!#1])-
c=[nX2]c=2-c-1=O
      probability: 0.9999785
    }
    tautomer{
      name: 3H,2-amino,6-one
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX3](-[ #1])-c=2[nX3](-[!#1])-
c=[nX2]c=2-c-1=O
      probability: 2.15e-05
    }
    tautomer{
      name: 2-amino,6-hydroxy
      pattern: [nX2]-1=c(-[NX3]-[ #1])-[nX2]=c-2[nX3](-[!#1])-c=[nX2]c-2=c-1-
O-[ #1]
      probability: 3.07e-06
    }
    tautomer{
      name: 1H,3H,2-imino,6-one
      pattern: [nX3]-1(-[ #1])-c(=[NX2])-[nX3](-[ #1])-c=2[nX3](-[!#1])-
c=[nX2]c=2-c-1=O
      probability: 2.80e-8
    }
}

tautomer_set{
    name: uracil-thymine
# Here we use (not counting H's and the N side-chain):
#               O
#               ||
#               C
#             /   \
#           N     C
#           |     |
#           C     C
#         // \   /
#        O     N
#
#
#
#               O2
#               ||
#               4
#             /   \
#           3     5
#           |     |
#           2     6
#         // \   /
#        O1    1


    tautomer{
      name: 1H,3H,2,4-dione
      pattern: n-1(-[ #1])-c(=O)-n(-[ #1])-c(=O)-c=c-1
      probability: 0.999999707
    }

    tautomer{
```

```
      name: 1H,2-hydroxy,4-one
      pattern: n-1(-[ #1])-c(-O-[ #1])=[nX2]-c(=O)-c=c-1
      probability: 1.87e-10
    }

    tautomer{
      name: 1H,2-one,4-hydroxy
      pattern: n-1(-[ #1])-c(=O)-[nX2]=c(-O-[ #1])-c=c-1
      probability: 2.93e-7
    }

    tautomer{
      name: 3H,2-hydroxy,4-one
      pattern: [nX2]-1=c(-O-[ #1])-n(-[ #1])-c(=O)-c=c-1
      probability: 4.59e-9
    }

    tautomer{
      name: 2-hydroxy,4-hydroxy
      pattern: [nX2]-1=c(-O-[ #1])-[nX2]=c(-O-[ #1])-c=c-1
      probability: 4.72e-12
    }

    tautomer{
      name: 3H,2-one,4-hydroxy
      pattern: [nX2]=1-c(=O)-n(-[ #1])-c(-O-[ #1])=c-c=1
      probability: 1.09e-09
    }

}

tautomer_set{
    name: uridine-thymidine
# Actually a methyl is used in place of the sugar which is OK for
# tautomers.
# This is similar to uracil except that there is one less H atom to move
# around and position 1 is blocked by the methyl

    tautomer{
      name: 3H,2,4-dione
      pattern: n1(-[!#1])-c(=O)-n(-[ #1])-c(=O)cc1
      probability: 0.99999724
    }

    tautomer{
      name: 2-hydroxy,4-one
      pattern: n1(-[!#1])-c(=O)-[nX2]=c(-O-[ #1])cc1
      probability: 2.76e-6
    }

    tautomer{
      name: 2-one,4-hydroxy
      pattern: n1(-[!#1])-c(-O-[ #1])=[nX2]-c(=O)cc1
      probability: 1.87e-10
    }
}
```

```
tautomer_set{
    name: cytosine
#Is this complete what about 1-3?
# Here we use (not counting H's and the N side-chain):
#              NH2
#              |
#              C
#            /   \
#           N     C
#           |     |
#           C     C
#         // \   /
#        O     N
#
#
#
#              N
#              |
#              4
#            /   \
#           3     5
#           |     |
#           2     6
#         // \   /
#        O     1

    tautomer{
      name: 1H,2-one,4-amino
      pattern: n-1(-[ #1])-c(=O)-[nX2]=c(-[NX3]-[ #1])-c=c-1
      probability: 0.996
    }

    tautomer{
      name: 3H,2-one,4-amino
      pattern: [nX2]=1-c(=O)-n(-[ #1])-c(-[NX3]-[ #1])=c-c=1
      probability: 0.00398
    }

    tautomer{
      name: 2-hydroxy,4-amino
      pattern: [nX2]-1=c(-O-[ #1])-[nX2]=c(-[NX3]-[ #1])-c=c-1
      probability: 5.63e-7
    }

    tautomer{
      name: 1H,3H,2-one,4-imino
      pattern: n-1(-[ #1])-c(=O)-n(-[ #1])-c(=[NX2])-c=c-1
      probability: 3.056e-6
    }

}

tautomer_set{
    name: 2,4-diaminopyrimidine
# (based on cytosine model)
#Is this complete what about 1-3?
# Here we use (not counting H's and the N side-chain):
```

```
#               NH2
#               |
#               C
#             /   \
#           N      C
#           |      |
#           C      C
#          / \    /
#        NH2   N
#
#
#
    tautomer{
      name: 1H,2-one,4-amino
      pattern: n-1(-[ #1])c(-[NX3])-[nX2]=c(-[NX3])-c=c-1
      probability: 0.5
    }

    tautomer{
      name: 3H,2-one,4-amino
      pattern: [nX2]-1=c(-[NX3])-n(-[ #1])=c(-[NX3])-c=c-1
      probability: 0.5
    }
}


tautomer_set{
    name: cytidine

#probabilities based upon corresponding structures for
#cytosine
    tautomer{
      name: 2-one,4-amino
      pattern: n1(-[!#1])c(=O)-[nX2]=c(-[NX3]-[ #1])cc1
      probability: 0.997
    }

    tautomer{
      name: 3H,2-one,4-imino
      pattern: n1(-[!#1])c(=O)-n(-[ #1])-c(=[NX2])cc1
      probability: 3.056e-6
    }

}

tautomer_set{
    name: adenine

    tautomer{
      name: 9H,6-amino
      pattern: [nX2]=1-c=[nX2]c=2-[nX3](-[ #1])-c=[nX2]c=2-c=1-[NX3]-[ #1]
      probability: 0.911
    }

    tautomer{
      name: 7H,6-amino
      pattern: [nX2]=1-c=[nX2]c=2-[nX2]=c-[nX3](-[ #1])c=2-c=1-[NX3]-[ #1]
```

```
      probability: 0.084
    }

    tautomer{
      name: 3H,6-amino
      pattern: [nX2]-1=c-[nX3](-[ #1])c-2=[nX2]-c=[nX2]c-2=c-1-[NX3]-[ #1]
      probability: 0.0046
    }

    tautomer{
      name: 1H,6-amino
      pattern: [nX3]-1(-[ #1])-c=[nX2]c-2=[nX2]-c=[nX2]c-2=c-1-[NX3]-[ #1]
      probability: 0.00043
    }

    tautomer{
      name: 1H,7H,6-imino
      pattern: [nX3]-1(-[ #1])-c=[nX2]c=2-[nX2]=c-[nX3](-[ #1])c=2-c-1=[NX2]
      probability: 2.88e-7
    }

    tautomer{
      name: 1H,9H,6-imino
      pattern: [nX3]-1(-[ #1])-c=[nX2]c=2-[nX3](-[ #1])-c=[nX2]c=2-c-1=[NX2]
      probability: 1.01e-6
    }

    tautomer{
      name: 3H,7H,6-imino
      pattern: [nX2]-1=c-[nX3](-[ #1])c=2-[nX3](-[ #1])-c=[nX2]c=2-c-1=[NX2]
      probability: 8.75e-10
    }

    tautomer{
      name: 3H,9H,6-imino
      pattern: [nX2]-1=c-[nX3](-[ #1])c=2-[nX2]=c-[nX3](-[ #1])c=2-c-1=[NX2]
      probability: 7.16e-12
    }
}

tautomer_set{
    name: adenosine

# based upon adenine probabilities
    tautomer{
      name: 6-amino
      pattern: [nX2]=1-c=[nX2]c=2-[nX3](-[!#1])-c=[nX2]c=2-c=1-[NX3]-[ #1]
      probability: 0.9999989
    }
    tautomer{
      name: 1H,6-imino
      pattern: [nX3]-1(-[ #1])-c=[nX2]c=2-[nX3](-[!#1])-c=[nX2]c=2-c-1=[NX2]
      probability: 1.1e-6
    }
    tautomer{
      name: 3H,6-imino
      pattern: [nX2]-1=c-[nX3](-[ #1])c=2-[nX3](-[!#1])-c=[nX2]c=2-c-1=[NX2]
```

```
      probability: 7.8e-12
    }

}
```