Proposed HELM RNA polymer nomenclature recommendations

Eric Swayze, Ionis Pharmaceuticals

Draft v4, 15 April, 2016

**Objective**

The goal of these recommendations is to provide a common set of guidelines for naming HELM monomers in an RNA polymer monomer database. The intent is to obtain revision and agreement from the alliance, and then publish the recommendations for all to adopt and draw from as is appropriate for their organization. Ionis intends to adopt the recommendations, name their monomer library accordingly, and deposit their publicly disclosed monomer library into the public domain.

**Sources**

These recommendations were derived from 2 sources:

1. IUPAC rules on biochemical nomenclature, Abbreviations and Symbols for Nucleic Acids, Polynucleotides and their Constituents, which are described in the 1970 recommendations at <http://www.chem.qmul.ac.uk/iupac/misc/naabb.html> . IUPAC abbreviation guidelines were chosen as a starting point due to their general use in the nucleic acid community, and the broad acceptance of IUPAC as the standard of chemical nomenclature and abbreviations.
2. The RNA Modification Database, which has abbreviations for a wide variety of modified RNA modifications, and generally uses an adaptation of the IUPAC recommendations. The online database is available <http://mods.rna.albany.edu/mods/> , and is further described in a publication <http://www.ncbi.nlm.nih.gov/pubmed/21071406> .

**General Concept**

The above mentioned systems abbreviate nucleosides, and provide ways to specify linkers. However, sugar modifications are indicated as modifications to the parent ribonucleoside. HELM requires sugars to be independent monomers. As such, the approach of the nomenclature system is to provide abbreviations for a given nucleoside’s base and sugar parts such that when recombined they provide a nucleoside Abbreviation that would be consistent with IUPAC recommendations for the common nucleosides.

**Recommendations**

1. Natural branch monomers (bases) are represented by capital letters A, C, G, T, U for adenine, cytosine, guanine, thymine, and uracil, respectively. Attachment points are those in the natural RNA nucleosides.
2. Modified branch monomers (bases) are indicated by a prefix to the natural monomer. Small letters are preferred. The position of the attachment should be indicated after the modification and before the natural monomer designation. Thus, 5-methylcytosine is m5C. 2-thiouracil is s2U. If a modification is used multiple times at the same position, the position is indicated first, for example, m62A for *N6,N6*-dimethyladenine. The use of sub and superscripts m62A would help visually, but is not compatible with database names. Modification codes are taken initially from IUPAC section [N-4.1](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p41), then the RNA modification database, then from other sources.
3. Backbone linkage monomers have phosphate as the natural analog. Phosphate is represented as lower case p. Modifications are named in a similar manner as for bases. Hence, phosphorothioate would be sp.
4. Backbone sugar monomers all have ribose as the natural analog. Ribose is represented as lower case r (lowercase is preferred as it is recommended by IUPAC and is the custom in RNA modification databases). Deoxyribose and 2’-O-methyl ribose are represented as d, and m, respectively. This keeps convention with IUPAC nomenclature where Am and Ad are 2’-O-methyladenosine and 2’-deoxyadenosine, respectively. Other modified sugars are represented with appropriate abbreviations in small letters. Where possible, modifications to sugars will follow the same abbreviation conventions as modifications to bases. For example, 4’-thioribose is s4r, and 2’-O-propylribose is pr2r.
5. Stereochemistry should be indicated with a capital letter R or S. If one stereochemistry is commonly used, omit the stereochemistry. For example, (S)-cEt is commonly used by Ionis, and is just cet, which is what it is called in papers, regulatory filings, etc. We also have the (R)-cEt version in research oligonucleotides, which is Rcet.
6. Difficult to name monomers will be assigned a code based on the chemical nomenclature if possible and reasonable. A complex monomer where this is unreasonable should be assigned a name as indicated in the first publication referencing it, or by a common name that is broadly accepted in the chemical literature. Examples of this are: moe for 2’-O-methoxyethylribose, lna for locked nucleic acid sugar, hna for hexitol nucleic acid sugar, cet for constrained (S)-2’-O-ethylribose, etc.
7. RNA polymer monomers provided by common vendors should be named in a manner similar to the abbreviation code of the preferred vendor, and have the vendor in the ‘name’ field. For example, a 1’,2’-Dideoxyribose (dSpacer) modification from IDT is abbreviated dSp (<https://www.idtdna.com/site/Catalog/Modifications/Product/1202> ). The RNA monomer would have a monmername (abbreviation) of dSp, and a name that has a more traditional name, and the vendor and associated code: 1',2'-Dideoxyribose (IDT dSpacer dSp). This provides one with a way to track the source if needed. Note that IDT includes an end code (5’, i for internal, 3’) but it is not needed so the recommended HELM nomenclature drops this addition.

**General Usage and Detailed Explanation of Rationale**

IUPAC recommendations have 1 letter codes to abbreviate nucleic acid monomers, but starting from the ribonucleoside and the phosphate linkage. For example, A is adenosine, pA is 5’-Adenosine monophosphate, ApG is an adenosine-guanosine dinucleotide. Modifications to the base nucleoside are indicated with a small letter after the nucleoside, for example Am is 2’-O-methyladenosine.

IUPAC provides a firm basis for abbreviating phosphate to p, and sequences to the natural nucleoside capital letter codes A, C, G, T, U. HELM requires further breaking the nucleoside into bases and sugars. However, the IUPAC nomenclature can be adapted to support this by using the capital letter to represent sequence (the base of the associated nucleoside), and the nucleoside modification code to represent the sugar.

As such, in the proposed HELM system, Am represents 2’-O-methyladenosine, just like IUPAC recommended abbreviation. However, structurally in the HELM system A represents the adenine monomer, and m represents 2’-O-methylribose sugar. To be consistent, r is the ribose sugar, and the natural analog for all sugars. Ar would then be a representation of adenosine using the HELM code system.

This also fits with the IUPAC system for modified bases of nucleosides, which employs modification codes as a prefix to the nucleoside. For example, m5C is 5-methylcytidine in IUPAC. The HELM system would use m5C for 5-methylcytosine. The IUPAC abbreviation for 2-thio-2’-O-methyluridine is s2Um. The HELM recommendation would be the same, with s2U meaning 2-thiouracil, and m representing 2’-O-methylribose.

**Usage derived from IUPAC sections**

1. [*N-3*](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p3)*. One-Letter Symbols.* This was the primary source for assigning key base names, which were adapted from the ribonucleoside designation according to the corresponding base on that ribonucleoside.
2. [*N-3.3*](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p33)*. Oligo- and Polynucleotides.* This describes notation for oligonucleotides and sugar modifications, the latter of which is indicated with a small letter indicating the modification after the bases.
3. [*N-4*](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p4)*. Modified Bases, Sugars, or Phosphates in Polynucleotides.* This lays the foundation for naming base modifications. Note most modifications to bases and sugars are indicated with small letters. Further, it specifies that base modifications should precede the nucleoside letter designation (meaning the base), and sugar modifications should be designated after the nucleoside designation.

**Specific Modifications and Adaptations from IUPAC**

1. In IUPAC recommendations for one-letter symbols, a dash or small p is used to designate a phosphate. To be consistent with single letter codes for natural monomers, p is used for phosphate. While modified linkages were not anticipated, modified linkages can easily be substituted for p. The inclusion of the p for phosphate is optional in IUPAC, but for HELM must be forced. ApG is an A-G dinucleotide.
2. IUPAC section [N-4.1](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p41) utilizes fl and io for fluoro and iodine, respectively, instead of the atomic symbols. This is done for consistency with natural RNA modifications containing formyl and isopentenyl groups which have been abbreviated historicaly as f and i, respectively. For consistency, we choose to keep this convention, though using fl2r for 2’-fluororibose is a bit cumbersome. If we agree, we can adopt f or fr for 2’-fluororibose.
3. Alkyl chains can be indicated by their length and the atom symbols of substitution points. Thus, an aminohexanol spacer is nC6o.
4. IUPAC rules and the RNA modification database both assume a ribose sugar. This is not consistent with the HELM system of using bases, sugars, and linkages to represent a nucleotide unit. The IUPAC nomenclature was adapted to employ the capital letter representing the nucleoside to represent the base of the associated nucleoside. IUPAC then specifies a modification code (d for deoxyribose, m for 2’-O-methylribose) that is placed after the nucleoside. Thus, Ad represents deoxyadenosine. IUPAC considers the d a deoxy modification to the adenosine nucleoside, while in HELM d is considered a deoxyribose sugar attached to an adenine base. Thus, the only modification from the IUPAC nomenclature at the nucleoside level would be a forced inclusion of r to represent a ribose nucleoside – Ar instead of A for adenosine. Likewise, IUPAC technically uses A to represent adenosine and Am is 2’-O-methyl-adenosine, with m representing the modification. In the recommended HELM system, A represents adenine, and m represents 2’-O-methylribose. This provides needed flexibility with extensively modified sugars, and fits with the HELM structural system while providing nearly identical nomenclature of the nucleoside unit.

**Appendix 1: Proposed RNA Monomer Names Derived From These Rules**

**RNA Branch Monomer (Base) Abbreviation Codes**

|  |  |  |
| --- | --- | --- |
| **Monomer Name** | **Natural Analog** | **Name** |
| A | A | Adenine |
| n2A | A | 2,6-Diaminopurine |
| bn6A | A | N6-benzyl-adenine |
| tclampA | A | T-clamp OMe |
| cpm6A | A | N-cyclopropylmethyl-adenine |
| m62A | A | N,N-dimethyl-Adenine |
| n2br7c7z8A | A | 7-deaza-8-aza-7-bromo-2-amino-Adenine |
| e6A | A | N6-ethyladenine |
| m6A | A | N6-Methyladenine |
| C | C | Cytosine |
| m5C | C | 5-methylcytosine |
| gclampC | C | G-Clamp (9-(aminoethoxy)phenoxazine) |
| br5C | C | 5-bromocytosine |
| z6C | C | 6-azacytosine |
| pry5C | C | 5-Propynyl-cytosine |
| G | G | Guanine |
| T | T | Thymine |
| m6T | T | 5,6-dimethyluracil |
| s2T | T | 2-thiothymine |
| z6T | T | 6-azathymine |
| U | U | Uracil |
| pry5U | U | 5-propyny-uracil |
| io5U | U | 5-iodouracil |
| fl5U | U | 5-fluorouracil |
| br5U | U | 5-bromouracil |
| m6U | U | 6-methyluracil |
| z6U | U | 6-azauracil |
| mo5U | U | 5-methoxyuracil |
| psiU | U | pseudouracil |
| Hyp | X | Hypoxanthine |
| fl24Ph | X | 2,4-Difluorophenyl |

**RNA Monomer (Sugar) Abbreviation Codes**

|  |  |  |
| --- | --- | --- |
| **Monomer Name** | **Natural Analog** | **Name** |
| r | r | ribose |
| d | r | deoxyribose |
| m | r | 2'-O-methylribose |
| moe | r | 2'-O-methoxyethyl ribose |
| fl2r, fr, f | r | 2'-fluororibose |
| pr2r | r | 2'-O-propylribose |
| npr2r | r | 2'-O-aminopropyl ribose |
| nma | r | 2'-O-(N-methylacetamide) ribose |
| lna | r | LNA (2'-O,4'-methylene bridged ribose) |
| mph | r | morpholino |
| fana | r | 2'-fluoroarabinose |
| s4r | r | 4’-thioribose |
| s4d | r | 4’-thio-2’-deoxyribose |
| 25r | r | (2'-5') Ribose |
| alna | r | alpha-L-LNA |
| m5d | r | 5'-methyldeoxyribose |
| ena | r | ENA (2'-O,4'-ethylene bridged ribose) |
| Rm5d | r | (R)-5'-methyldeoxyribose |
| Sm5d | r | (S)-5'-methyldeoxyribose |
| Rcet | r | (R)-cEt BNA |
| cet | r | (S)-cEt BNA |
| Rm5fl2r | r | (5'R)-2'-fluoro-5'-methyl-ribose |
| Sm5fl2r | r | (5'S)-2'-fluoro-5'-methyl-ribose |
| Scmoe | r | (S)-cMOE BNA |
| Rcmoe | r | (R)-cMOE BNA |
| tcdna | r | tricyclo DNA Sugar |
| fhna | r | 3'-fluoro-HNA |
| mana | r | 3'-O-Methyl-ANA |
| Rmclna | r | 2'-(R)-methyl-cLNA |
| Smclna | r | 2'-(S)-methyl-cLNA |
| cipr | r | ciPr BNA |
| hna | r | Hexitol Nucleic Acid |
| mclna | r | methylene cLNA |
| almclna | r | alpha-L-methylene cLNA |
| s2lna | r | 2'-Thio LNA |
| Rflclna | r | (R)-F-cLNA |
| afhna | r | 3'-ara-FHNA |
| fleana | r | 3'-O-FEt ANA |
| fcena | r | Fluoro-CeNA |
| una | r | UNA (2'-3'-Unlocked-ribose) |
| 253d | r | (2'-5') 3'-deoxyribose |
| cena | r | CeNA |
| dSp | r | 1',2'-Dideoxyribose (IDT dSpacer dSp) |
| Rgna | r | R Propanetriol (GNA sugar) |
| Sgna | r | S Propanetriol (GNA sugar) |
| n3r | r | 3-aminoribose |
| nC12o | r | 12-aminododecanol |
| nC6o | r | 6-aminohexanol |

**RNA Monomer (Linker) Abbreviation Codes**

|  |  |  |
| --- | --- | --- |
| **Monomer Name** | **Natural Analog** | **Name** |
| p | p | phosphate |
| sp | p | phosporothioate |
| mp | p | methylphosphonate |
| m2np | p | phosphorodiamidate |
| Ssp | p | (Sp)-phosporothioate |
| Rsp | p | (Rp)-phosporothioate |
| s2p | p | phosphorodithioate |
| bp | p | Boranophosphate |

**CHEM Monomer Abbreviation Codes**

|  |  |
| --- | --- |
| **Monomer Name** | **Name** |
| 36SS | 3'-Thiol-Modifier 6 S-S (Glen Research) |
| 3Biop | 3' Biotin PO (IDT 3Bio) |
| 3Bios | 3' Biotin PS (IDT 3Bio) |
| 3FAM | 3'-FAM (IDT 36-FAM) |
| 5Bio2p | 5' Dual Biotin PO (IDT 52-Bio) |
| 5Bio2s | 5' Dual Biotin PS (IDT 52-Bio) |
| 5BioC6 | 5' Biotin aminohexyl (Glen 5'-Biotin Phosphoramidite 10-5950) |
| 5BioTEG | 5' Biotin TEG (IDT 5BiotinTEG) |
| 5CholTEG | 5' cholesterol-TEG (Glen 10-1976) |
| 5Cy3 | 5'-Cy3 (IDT 5Cy3) |
| 5Cy3nC6o | 5'-Cy3-hexylamino |
| 5Cy5 | 5'-Cy5 (IDT 5Cy5) |
| 5Cy55 | 5'-Cy5.5 (IDT) |
| 5DBCOTEG | 5' DBCO-TEG (Glen 10-1941) |
| 5FAM | 5'-FAM (IDT 56-FAM) |
| 5PCBio | 5' biotin photocleavable (IDT 5PCBio) |
| 5sf2Cy3nC6o | 5'-SulfoCy3-hexylamino |
| 5THAGN | 5'-THA-C6-GalNAc3 |
| BioTEG | Biotin TEG (Glen BiotinTEG Phosphoramidite 10-1955) |
| deSBioTEG | desthiobiotin TEG (IDT deSBioTEG) |
| FAMncs | Fluorescein thiourea phosphoramidite (Glen Research 10-1963-xx) |
| nC6o | 6-aminohexanol |
| hxyno | Hexynyl alcohol |
| PEG2 | Diethylene Glycol |
| sDBL | Symmetric Doubler from Glen Research (10-1920-xx) |
| SMCC | SMCC from ThermoFisher (succinimidyl 4-(N-maleimidomethyl)cyclohexane-1-carboxylate) |

**Appendix 2: Selected IUPAC Rule sections**

[**N-4.1**](http://www.chem.qmul.ac.uk/iupac/misc/naabb.html#p41)**. Designation of Substituents on Bases**

In long sequences, as in transfer RNA's, where it is preferable to have not more than one capital letter per nucleoside residue, the standard symbols for nucleosides [i.e., A, U, G, C, etc. (see N-3.2.1)] may be modified by a symbol of lower case letter(s) placed immediately before the single capital letter. Those symbols recommended for more common modifications are listed below (for locants and multipliers, see N-4.4; for unusual sugar residues, see N-3.2.2 and N-3.2.3:

|  |  |
| --- | --- |
| m, e, ac | methyl, ethyl, acetyl |
| n, o | amino (N replaces H), deamino (O replaces N) |
| z, c | aza (N replaces C), deaza (C replaces N) |
| h | dihydro (hU = dihydrouridine; see also N-3.2.1 and N-4.4) |
| hm, ho (or oh) | hydroxymethyl, hydroxy |
| aa | aminoacyl |
| f | formyl (as in the conventional fMet for formylmethionyl) |
| fa | formylaminoacyl |
| i | isopentenyl (= γ,γ-dimethylallyl) |
| s | thio or mercapto (sU = thiouridine; see also N-3.2.1 and N-4.4) |
| fl, cl, br, io | fluoro, chloro, bromo, iodo (not encountered in natural polynucleotides;  see also N-3.2.1 and N-4.4). |

Symbols for some N-protecting radicals used in synthetic work [4, 9] are:

|  |  |
| --- | --- |
| bz, bzl, tos | benzoyl, benzyl, tosyl |
| tr, an, bh | trityl, anisoyl, benzhydryl (diphenylmethyl) |
| mmt | monomethoxytrityl(*p*-anisyldiphenylmethyl) |
| dmt | dimethoxytrityl (di-*p*-anisylphenylmethyl) |
| thp, dns | tetrahydropyranyl, dansyl |
| cmc | *N*-cyclohexyl*-N'*[β-(4-methylmorpholino) amidino]  (reaction product from the corresponding carbodiimide) [16]. |