Eidogen-Sertanty

Content Prioritization And Content Entry and Quality Control Process

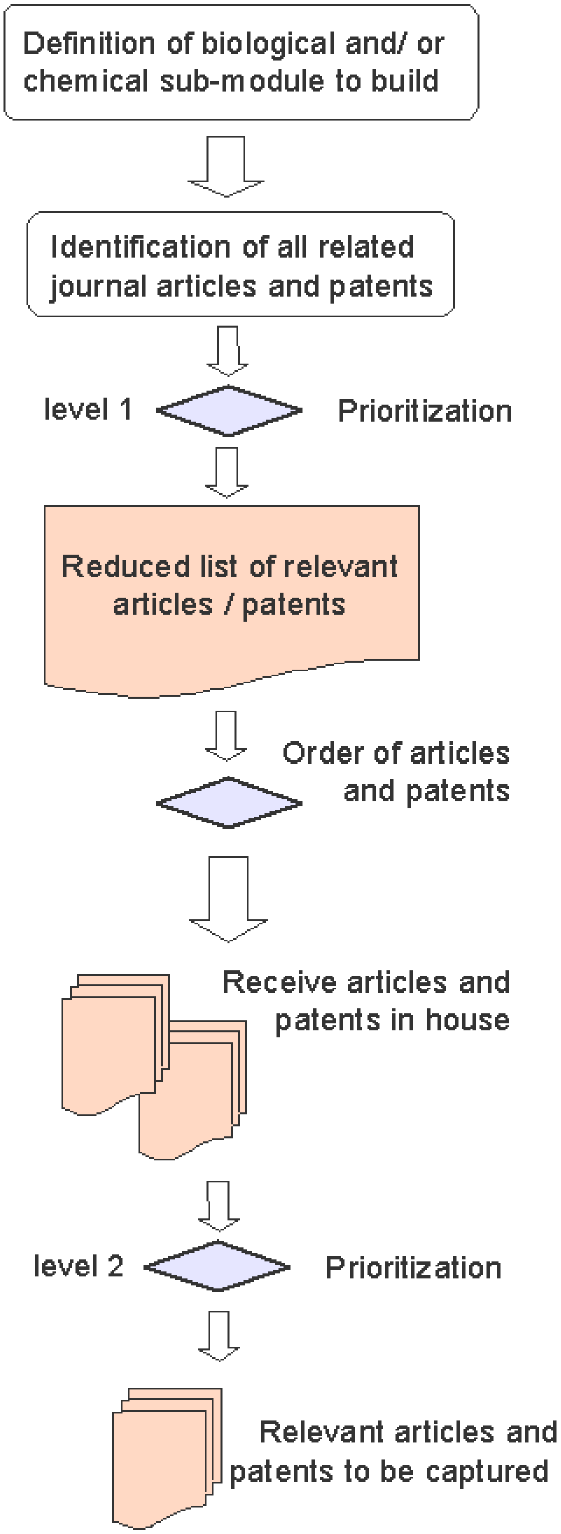
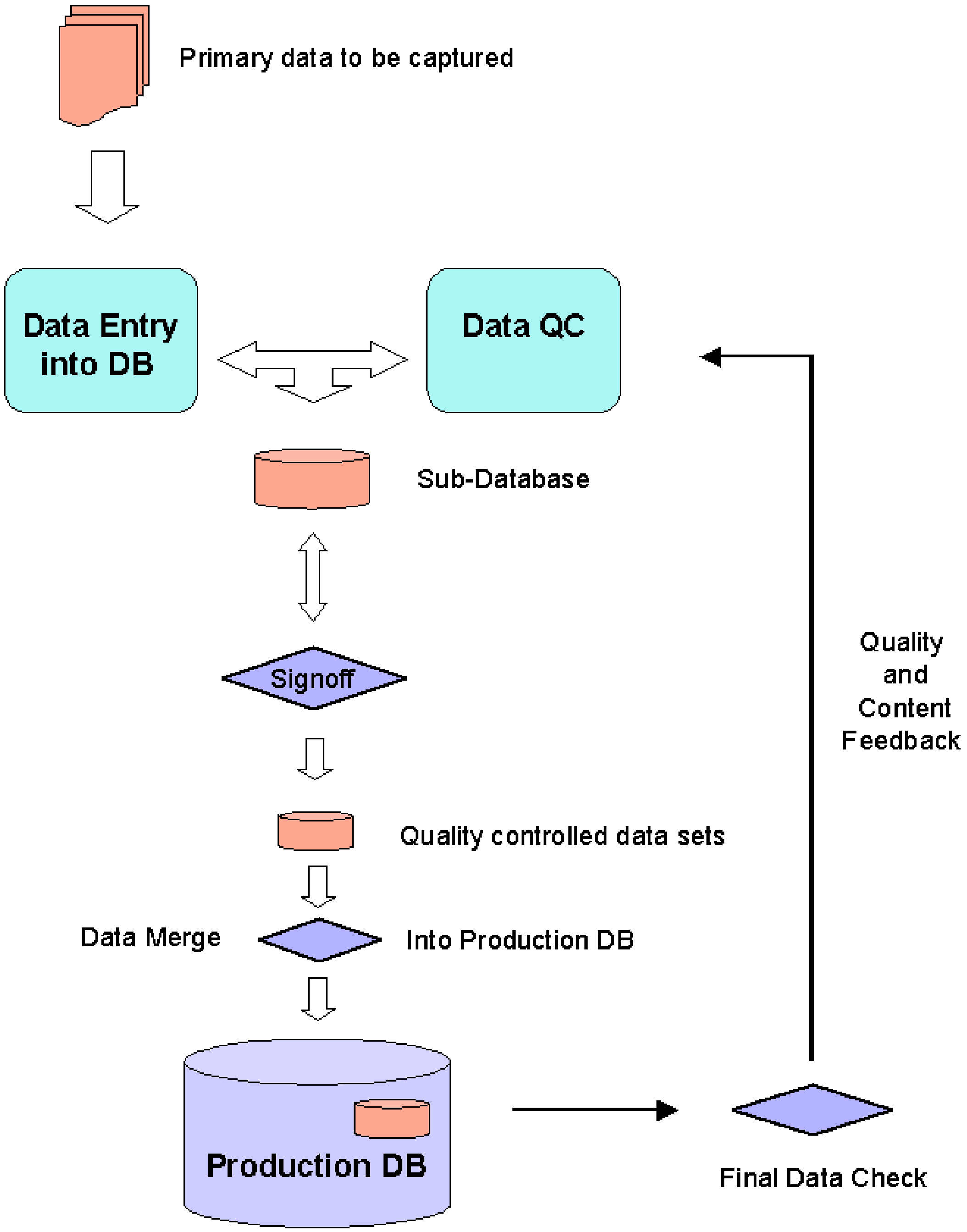

Definition of biological and/ or chemical sub-module to build
Identification of all related journal articles and patents
level 1
Prioritization
Reduced list of relevant articles / patents
Order of articles and patents
Receive articles and patents in house
level 2
Prioritization
Relevant articles and patents to be captured

Figure 1. Workflow of retrieval of relevant primary data for content development

The process of data capture begins with the definition of the content module or sub-module to be built (see figure 1). Broadly we define biological and chemistry space we want to capture and compile the key scientific terms to identify the related journal articles and patents. A number of searches in different public and patent databases are performed and a comprehensive list of related publications is generated. We have developed a lot of expertise to do these searches comprehensively. Scientists prioritize this list using title, keywords and abstract (if available). A reduced list of publications that most likely contain relevant data is generated. We use four categories (most relevant, relevant, less relevant, not relevant) and the most relevant articles and patents are ordered.

Obtained patents are further prioritized into 6 categories determining if and what biological activity data they contain and in which format the data is published. In-vitro enzyme activity data, in particular IC50 values are most relevant for the development of QSAR models (ePotency). Cell-based activity and toxicity data are relevant for development of advanced activity models and also eADME and eTox. Purely biologically oriented publications, e.g. focused on the analysis of complex cellular signaling pathways, protein analysis, synergy effects, etc. are prioritized as less important. Publications considered most relevant and relevant are captured in the Sertanty database. Categorization can depend on customer preferences.



\

Figure 2. Workflow of content capture; data entry, quality control, signoff

The data capture process (see figure 2) is a sequence of data entry and quality control (by different people). All structure-activity data points are associated with a comprehensive biological assay procedure (if provided) and in most cases categorized by a standardized target name (official approved gene symbol, see below), assay type, and (if applicable) cell line and species. Also we emphasize the scientific context of the information and ensure each scientist has a thorough understanding of the subject. Scientific context information is captured in the database as additional comments. A collection of structure-activity data points associated with a distinct assay procedure, target, etc. (as described above) is called a biological protocol. For each protocol targets are categorized by an ontology based on function and mechanism. For kinases we follow the SwissProt EC nomenclature (see below). We also capture reference and a Web-link (if available) to the primary data source.

For database entry remains associated with its creator and the date when it was generated. After an entry is marked as QC passed, only the creator or an administrator can modify this entry. If an article fails QC, it goes back to its creator for correction. QC-passed patents and articles are then signed off for quality and consistence of content by a team leader. All final articles that are signed off are merged to the main Sertanty production database. For each article or patent, all steps of data entry, data QC, and signoff are documented, associated with the individuals involved, dated, and signed. Every two weeks all units that are signed off are sent and merged into the production database. There, the entries are briefly revisited and checked for consistency, feedback is provided to the data entry/QC team. In rare cases corrections are made or data must be revisited by data entry/QC; an administrator had the right to reset a QC flag for correction or change of information. Data that passed the final check is ready to be shipped to customers.

Structure activity data points related to a specific target are then further analyzed by a medicinal chemists according to their experimental assay procedure, mechanism and mode of action, potential binding site to the target, etc. and are grouped together if the data points are comparable. These groups of SAR data points are then used to develop eScreen models.

Target Standardization and Classification

We devote particular effort in the consolidation, standardization, and classification of the kinase target names and eScreen groups. We use controlled vocabulary of approved symbols to achieve consistency and data integrity among the whole database. This ensures maximal value for the medicinal research scientists and for computational chemists who use the data to derive QSAR models.

For target names we use approved human gene symbols according to HUGO gene nomenclature committee[1], NCBI's Locus Link[2], and Expasy's Swiss Prot[3]. The same symbols are used for eScreen groupings. To build models we also include additional information in the grouping symbols as appropriate (e.g. _DAG for the diacylglycerol binding site of protein kinase C targets or SH2 for SRC-homology binding; the default binding site is the ATP binding region).

We classify kinase targets considering structural and functional similarity according to Hanks, an internationally accepted standard[4] and recently – as the Hanks classification is no longer updated – according to kinase classification developed by Sugen[5].

We also capture an exhaustive dictionary of target synonyms and update target symbols if the official symbol changes.

We consider controlled vocabulary and thorough classification of targets an important consideration in the data base production process, in order to develop a product that can help analyzing how structural and functional target similarity corresponds to similarity among targets based on the efficacy of small molecule inhibitors against different kinase targets. Eventually this will lead to understand the big picture of the extremely challenging field of discovery of potent and selective small molecule kinase inhibitors.

[1] http://www.gene.ucl.ac.uk/nomenclature/
[2] http://www.ncbi.nih.gov/LocusLink/
[3] http://us.expasy.org/sprot/
[4] http://pkr.sdsc.edu/html/pk_classification/pk_catalytic/pk_hanks_class.html
[5] http://198.202.68.14/human/kinome/

Table 1 below shows the classification of targets for which eScreen models have been developed categorization by family, group and binding site.

eScreen Name	Target Symbol	Binding Site	FAMILY	GROUP
PKA	PKA	ATP	PKA	AGC
PKC	PKC	ATP	PKC	AGC
PRKCA	PRKCA / PKCa	ATP	PKC	AGC
PRKCB1	PRKCB1 / PCKb	ATP	PKC	AGC
PRKCD	PRKCD / PCKd	ATP	PKC	AGC
PRKCA_DAG	PRKCA / PKCa	DAG	PKC	AGC
PRKCB1_DAG	PRKCB1 / PCKb	DAG	PKC	AGC
PRKCE_DAG	PRKCE / PKCe	DAG	PKC	AGC
PRKCG_DAG	PRKCG / PKCg	DAG	PKC	AGC
PRKCH_DAG	PRKCH / PKCh	DAG	PKC	AGC
PDK	PDK	ATP	PDK	AGC / atypical
CDC2	CDC2 / CDK1	ATP	CDK	CMGC
CDK1_B	CDC2 / CDK1	ATP	CDK	CMGC
CDK2	CDK2	ATP	CDK	CMGC
CDK2_A	CDK2	ATP	CDK	CMGC
CDK2_E	CDK2	ATP	CDK	CMGC
CDK4_D1	CDK4	ATP	CDK	CMGC
CDK4_D2	CDK4	ATP	CDK	CMGC
CDK5	CDK5	ATP	CDK	CMGC
CDK5_P25	CDK5	ATP	CDK	CMGC
CDK5_P35	CDK5	ATP	CDK	CMGC
MAPK14	MAPK14 / p38a	ATP	MAPK	CMGC
ADK	ADK	ATP	MAPK	CMGC
GSK3A	GSK3A	ATP	GSK	CMGC
GSK3B	GSK3B	ATP	GSK	CMGC
CSK	CSK / c-Src	ATP	Csk	TK-NR
SRC	SRC / v-Src	ATP	Src	TK-NR
FYN	FYN	ATP	Src	TK-NR
LCK	LCK	ATP	Src	TK-NR
ABL	ABL1	ATP	Abl	TK-NR
SYK	SYK	ATP	Syk	TK-NR
ZAP_70_SH2	ZAP70	SH2	Syk	TK-NR
GRB2_SH2	GRB2	SH2	Src / unknown	TK-NR / unknown

EGFR	EGFR	ATP	EGFR	TK-R
ERBB2	ERBB2 / HER2	ATP	EGFR	TK-R
FGFR	FGFR	ATP	FGFR	TK-R
EGFR_SUBSTRATE	EGFR	Substrate	FGFR	TK-R
PDGFR	PDGFRA	ATP	PDGFR	TK-R
PDGFRB	PDGFRB	ATP	PDGFR	TK-R
FLT1	FLT1	ATP	VEGFR	TK-R
KDR	KDR	ATP	VEGFR	TK-R
TEK	TEK / TIE2	ATP	Tie	TK-R
RAF1	RAF1	ATP	RAF	TKL
MAP2K	MAP2K1	ATP	STE-7	STE
PPK	PPK	ATP	unknown	unknown
PIK4CA	PIK4CA	ATP	Inositol kinase Family	unknown

Table 1. Classification of eScreen targets according to Sugen and by binding site

Table 2 shows an example for more detailed classification and synonyms for

Symbol in DB	Standard Gene Symbol	Gene Ontology	Group	Family	Sub-family	Synonyms
ABL1	ABL1	TK-NR	TK	ABL		ABL1 ABL JTK7 p150 c-ABL ABL p43 p43abl
CDK4	CDK4	ST-NR	CMCG	CDK	CDK4	CDK4 Cyclin-dependent kinase 4 PSK-J3 CDK4/D1 CDK4/D CDK4/D2 CDK4/D3
MAPKAPK2	MAPKAPK2	MAP	CAMK	MAPKAPK	MAPKAPK	MAPKAPK2 MAPK-activated protein kinase 2 MAPKAP kinase 2 MAPKAPK-2
MAPK14	MAPK14	MAP	CMCG	MAPK	P38	MAPK14 p38 CSBP p38alpha p38-alpha
PRKCA	PRKCA	ST-NR	AGC	PKC	ALPHA	PRKCA PKCA PRKACA PKC-alpha PKC-A PRKC
TEK	TEK	TK-R	TK	TIE		TEK TIE2 Tyrosine-protein kinase receptor TIE-2 Tyrosine-protein kinase receptor TEK P140 TEK Tunica interna endothelial cell kinase CD202b antigen
ZAP70	ZAP70	TK-NR	TK	SYK		ZAP70 SRK 70 kDa zeta-associated protein Syk-related tyrosine kinase STD p70zap ZAP ZAP-70
LCK	LCK	TK-NR	TK	SRC		LCK P56-LCK LSK T cell-specific protein-tyrosine kinase Tck(m) p53/p56lck Lck-SH2
ITK	ITK	TK-NR	TK	TEC		ITK LYK EMT T-cell-specific kinase Tyrosine-protein kinase Lyk Kinase EMT Tsk(m) PSCTK2

Table 2. Classification of selected kinase targets and synonyms