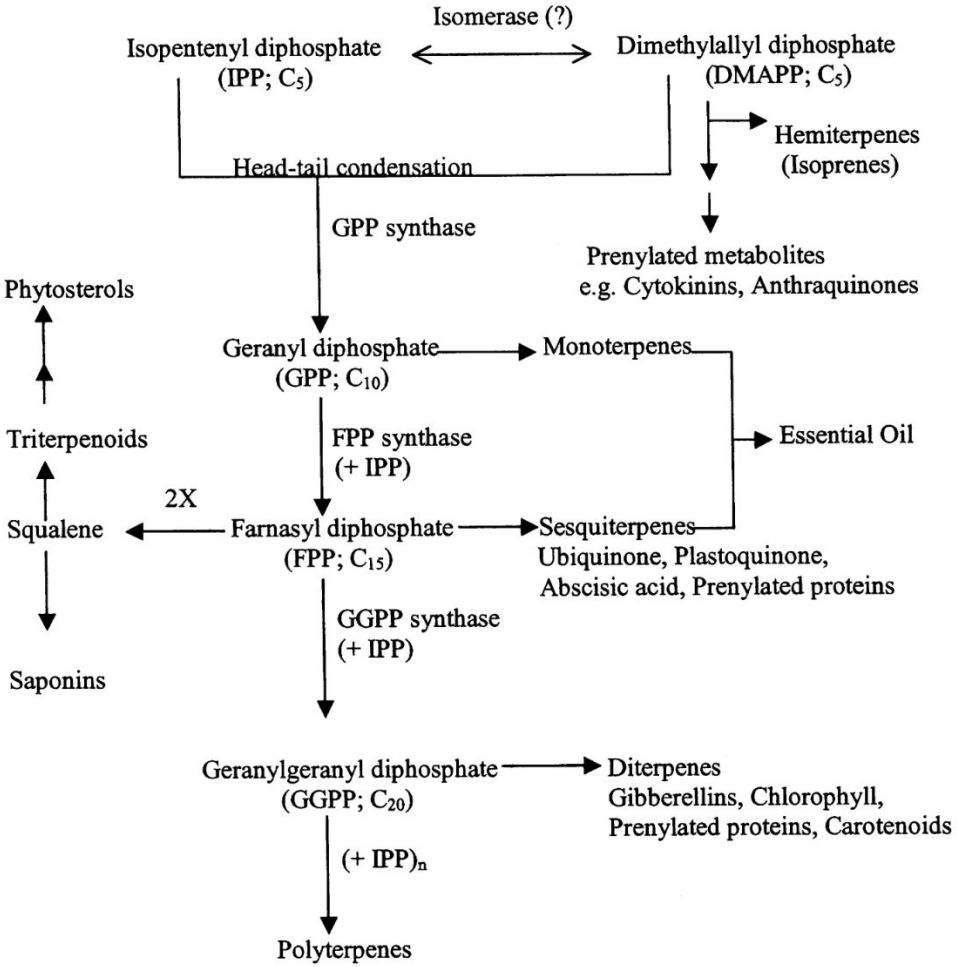
Biosynthetic pathway extraction from images

Plan

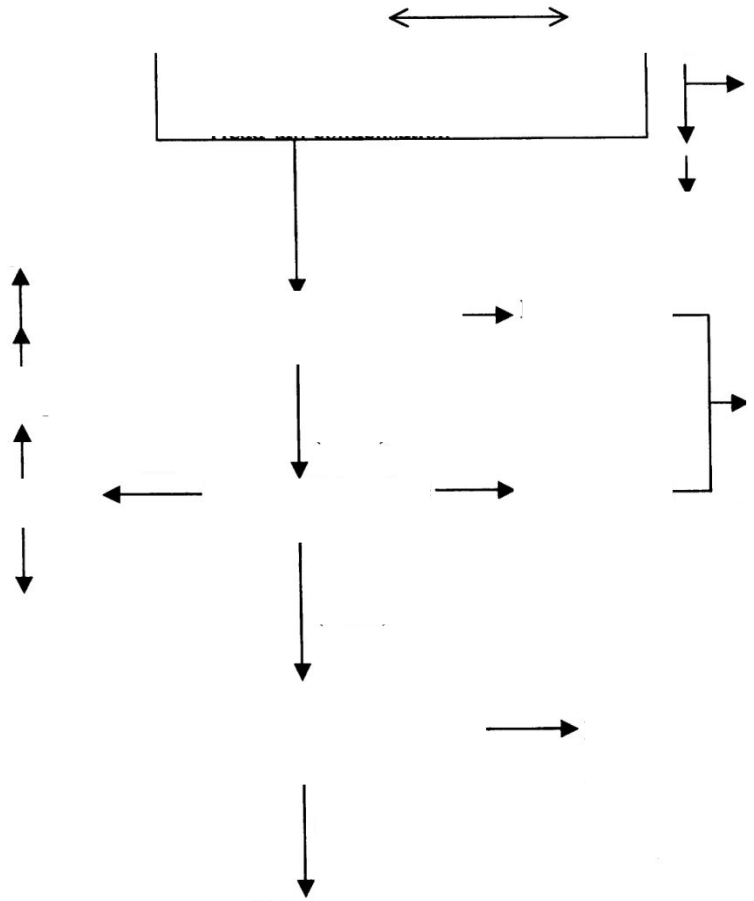
The problem can broadly be divided into two parts:

- Logical structure of the diagram
- The contextual meaning of the text

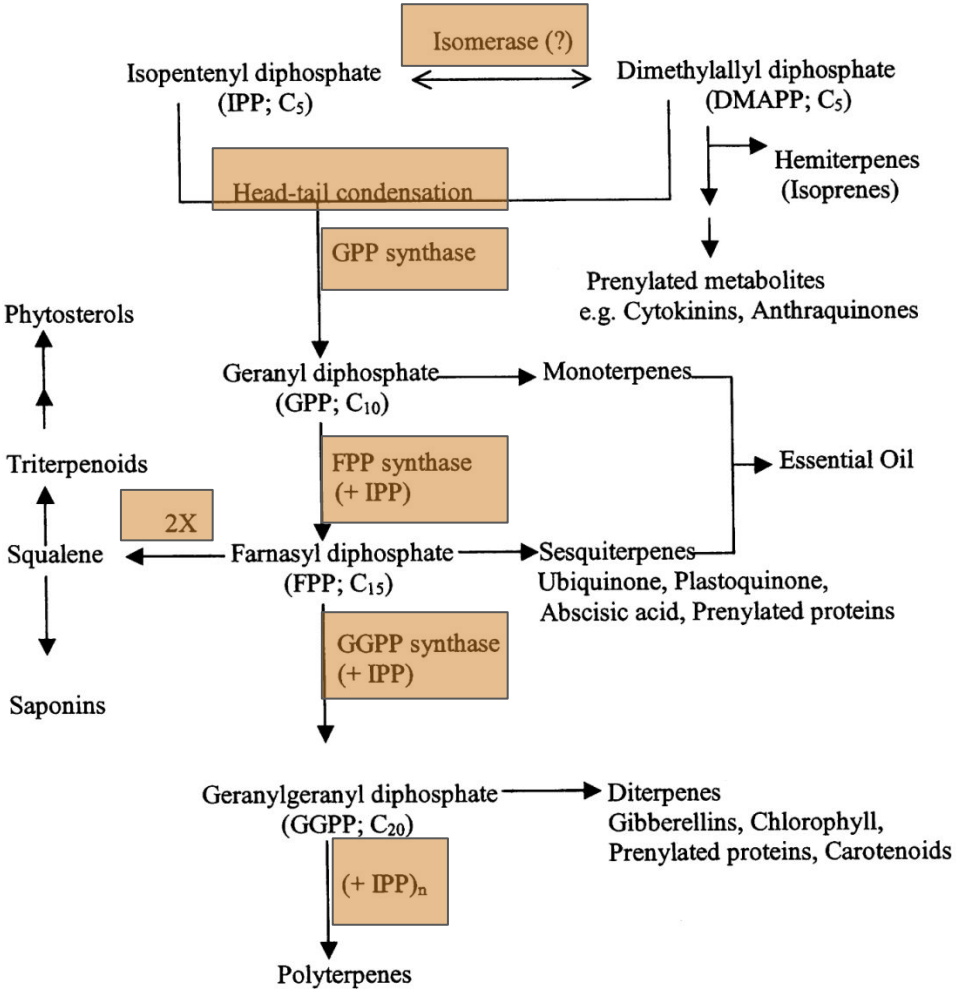
Original Image



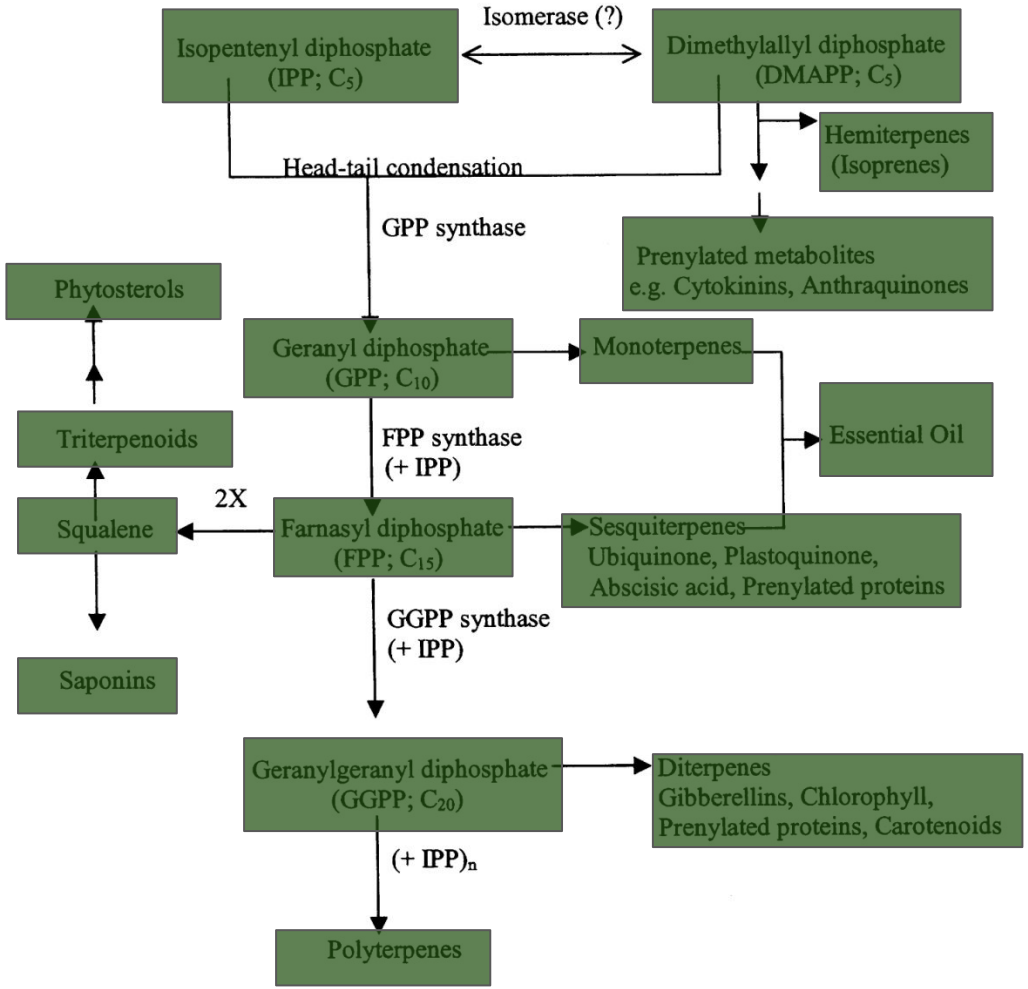
Arrow Layer



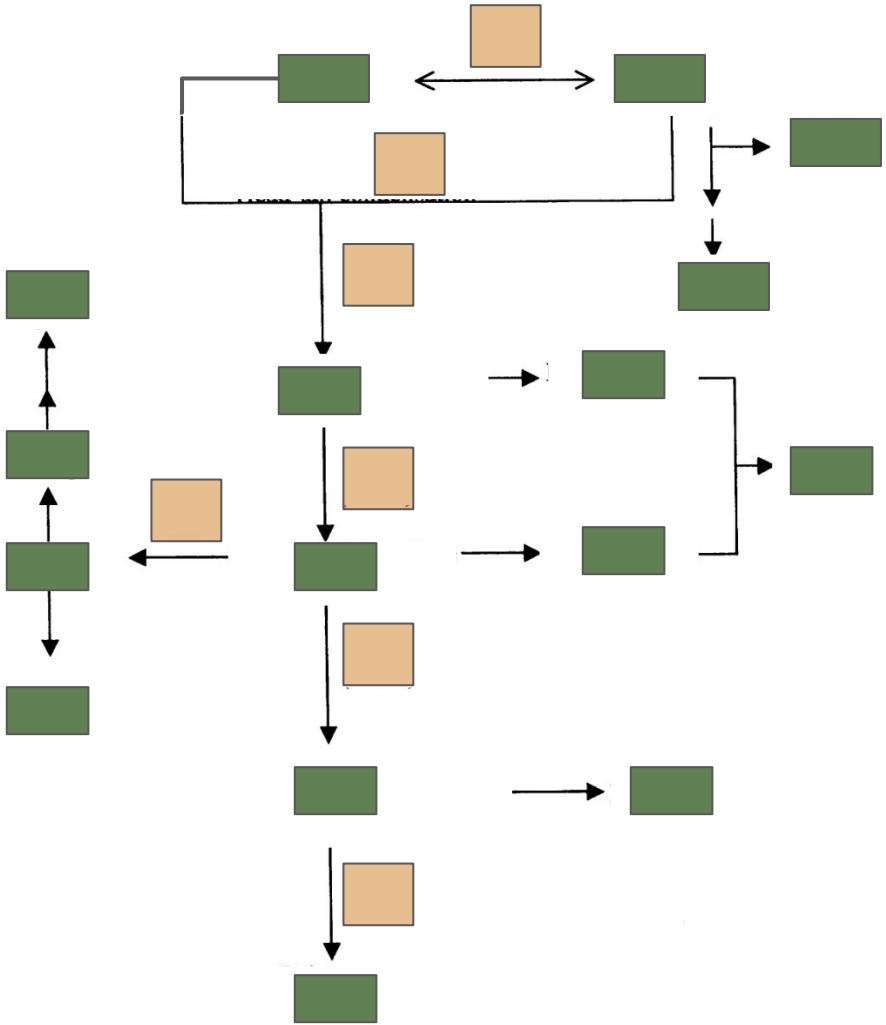
Image


Isomerase (?)
Isopentenyl diphosphate
(IPP; C5)
Dimethylallyl diphosphate
(DMAPP; C5)
Hemiterpenes
(Isoprenes)
Head-tail condensation
GPP synthase
Prenylated metabolites
e.g. Cytokinins, Anthraquinones
Phytosterols
Geranyl diphosphate
(GPP; C10)
Monoterpenes
Triterpenoids
FPP synthase
(+ IPP)
Essential Oil
2X
Squalene
Farnasyl diphosphate
(FPP; C15)
Sesquiterpenes
Ubiquinone, Plastoquinone,
Abscisic acid, Prenylated proteins
GGPP synthase
(+ IPP)
Saponins
Geranylgeranyl diphosphate
(GGPP; C20)
Diterpenes
Gibberellins, Chlorophyll,
Prenylated proteins, Carotenoids
(+ IPP)n
Polyterpenes

Products and Reactants Layer

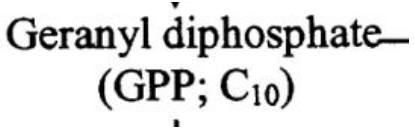
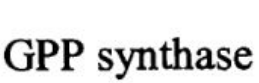
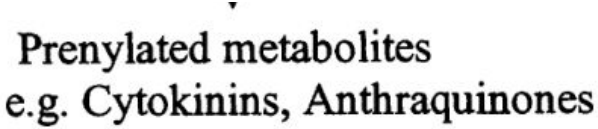


Logical
Semantics



Challenges

- Overlapping characters and arrows
- Curved arrows
- No standard format for biosynthetic pathway diagrams

Image segment		Text output
	Tesseract →	Geranyl diphosphate (GPP; C_{10})
	Tesseract →	GPP synthase
	Tesseract →	Prenylated metabolites e.g. Cytokinins, Anthraquinones

Geranyl diphosphate—
(GPP; C_{10})

Partial Chemical formula

Chemical Name → Geranyl diphosphate (GPP; C_{10})

Synonym/abbreviation

Prenylated metabolites
e.g. Cytokinins, Anthraquinones

Chemical Type → Prenylated metabolites
e.g. Cytokinins, Anthraquinones

Examples

Challenges

- Low resolution JPEGS
- Overlapping characters
- Different kinds of information present

Plan

The problem can broadly be divided into two parts:

- Logical structure of the diagram
- The contextual meaning of the text

Week 1 - 2 (Done):

- Project planning
- Selecting target images
- Exploring image processing tools

Week 3 - 4:

- Resolving arrows
- Segmenting the image
- OCR of Islands of text

Week 5 - 6:

- Creating diagram structure map from image segments and arrow orientations
- Annotating terms using dictionary