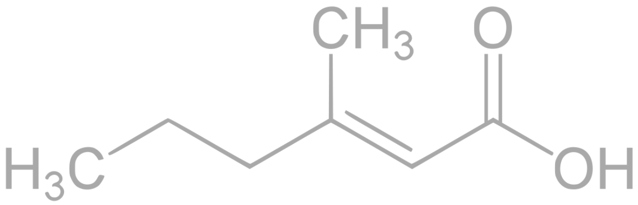
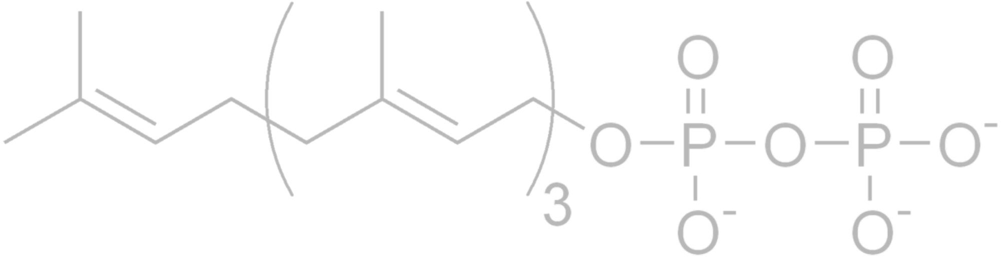


Extraction of Biosynthetic Pathways from Images

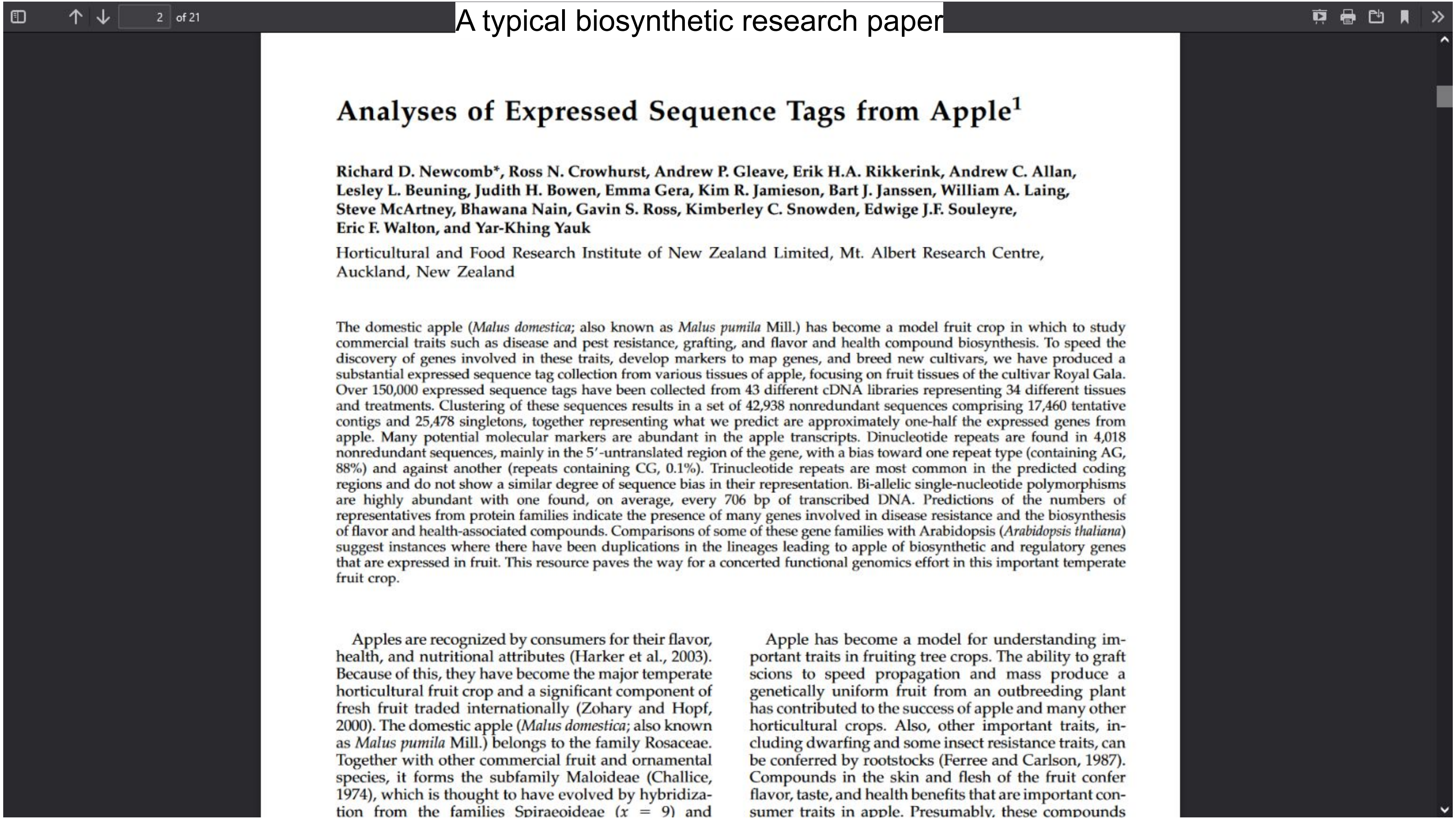
Anubhab Chakraborty



Overview

- A typical research paper of our interest
- Goals for pyamiimage
- Running Tesseract OCR on png image
- Generating SVG from Tesseract OCR
- Phrase detection
- Phrase annotation
- Contextual arrows for chemical reactions
- Extracting arrows from biosynthetic pathways
- Adding contextual arrows to SVG
- Where to find us

A typical biosynthetic research paper



Analyses of Expressed Sequence Tags from Apple[1]

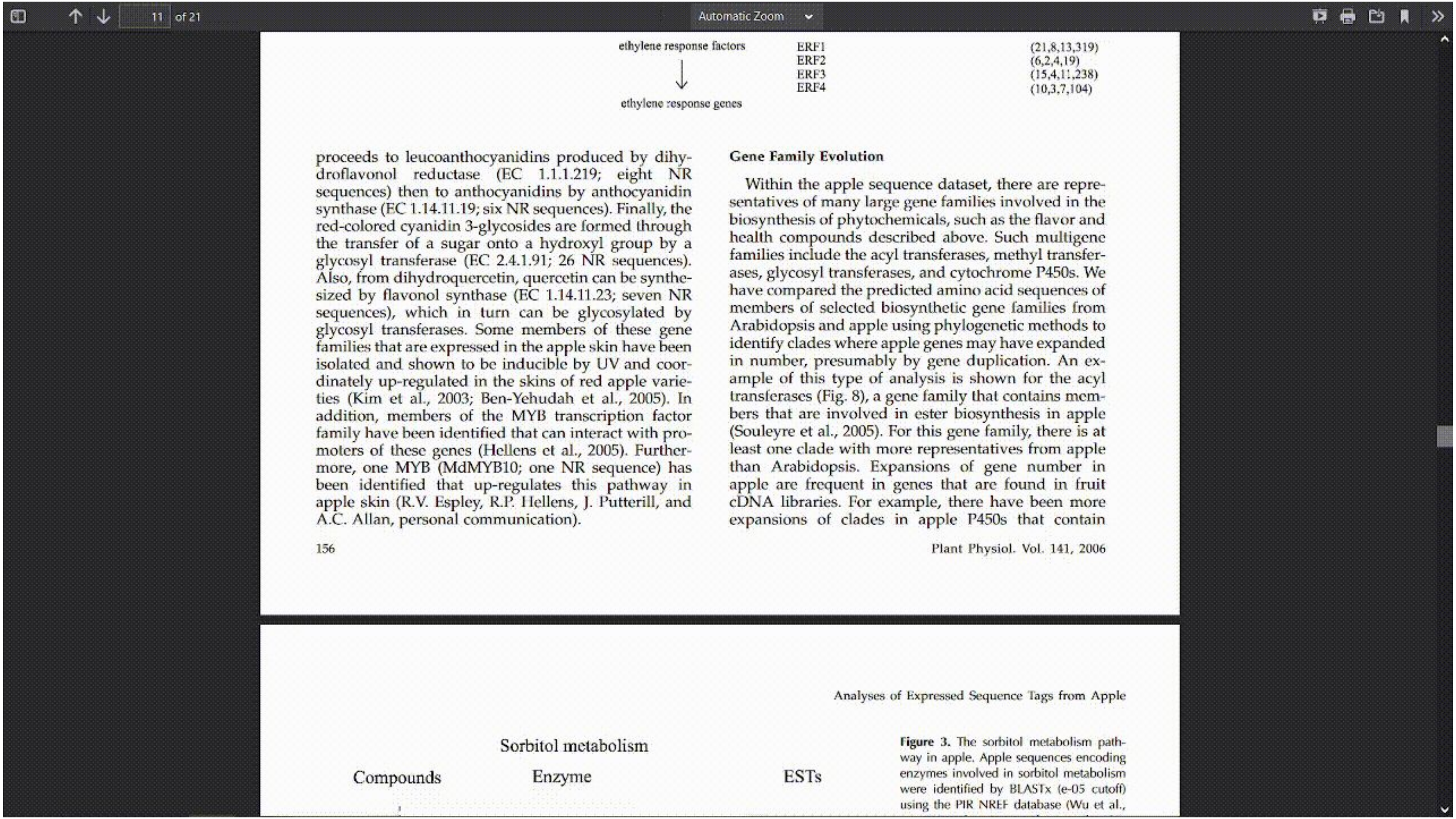
Richard D. Newcomb*, Ross N. Crowhurst, Andrew P. Gleave, Erik H.A. Rikkerink, Andrew C. Allan, Lesley L. Beuning, Judith H. Bowen, Emma Gera, Kim R. Jamieson, Bart J. Janssen, William A. Laing, Steve McArtney, Bhawana Nain, Gavin S. Ross, Kimberley C. Snowden, Edwige J.F. Souleyre, Eric F. Walton, and Yar-Khing Yauk

Horticultural and Food Research Institute of New Zealand Limited, Mt. Albert Research Centre, Auckland, New Zealand

The domestic apple (*Malus domestica*; also known as *Malus pumila* Mill.) has become a model fruit crop in which to study commercial traits such as disease and pest resistance, grafting, and flavor and health compound biosynthesis. To speed the discovery of genes involved in these traits, develop markers to map genes, and breed new cultivars, we have produced a substantial expressed sequence tag collection from various tissues of apple, focusing on fruit tissues of the cultivar Royal Gala. Over 150,000 expressed sequence tags have been collected from 43 different cDNA libraries representing 34 different tissues and treatments. Clustering of these sequences results in a set of 42,938 nonredundant sequences comprising 17,460 tentative contigs and 25,478 singletons, together representing what we predict are approximately one-half the expressed genes from apple. Many potential molecular markers are abundant in the apple transcripts. Dinucleotide repeats are found in 4,018 nonredundant sequences, mainly in the 5′-untranslated region of the gene, with a bias toward one repeat type (containing AG, 88%) and against another (repeats containing CG, 0.1%). Trinucleotide repeats are most common in the predicted coding regions and do not show a similar degree of sequence bias in their representation. Bi-allelic single-nucleotide polymorphisms are highly abundant with one found, on average, every 706 bp of transcribed DNA. Predictions of the numbers of representatives from protein families indicate the presence of many genes involved in disease resistance and the biosynthesis of flavor and health-associated compounds. Comparisons of some of these gene families with Arabidopsis (*Arabidopsis thaliana*) suggest instances where there have been duplications in the lineages leading to apple of biosynthetic and regulatory genes that are expressed in fruit. This resource paves the way for a concerted functional genomics effort in this important temperate fruit crop.

Apples are recognized by consumers for their flavor, health, and nutritional attributes (Harker et al., 2003). Because of this, they have become the major temperate horticultural fruit crop and a significant component of fresh fruit traded internationally (Zohary and Hopf, 2000). The domestic apple (*Malus domestica*; also known as *Malus pumila* Mill.) belongs to the family Rosaceae. Together with other commercial fruit and ornamental species, it forms the subfamily Maloideae (Challice, 1974), which is thought to have evolved by hybridization from the families Spiraeoideae ($x = 9$) and

Apple has become a model for understanding important traits in fruiting tree crops. The ability to graft scions to speed propagation and mass produce a genetically uniform fruit from an outbreeding plant has contributed to the success of apple and many other horticultural crops. Also, other important traits, including dwarfing and some insect resistance traits, can be conferred by rootstocks (Ferree and Carlson, 1987). Compounds in the skin and flesh of the fruit confer flavor, taste, and health benefits that are important consumer traits in apple. Presumably, these compounds

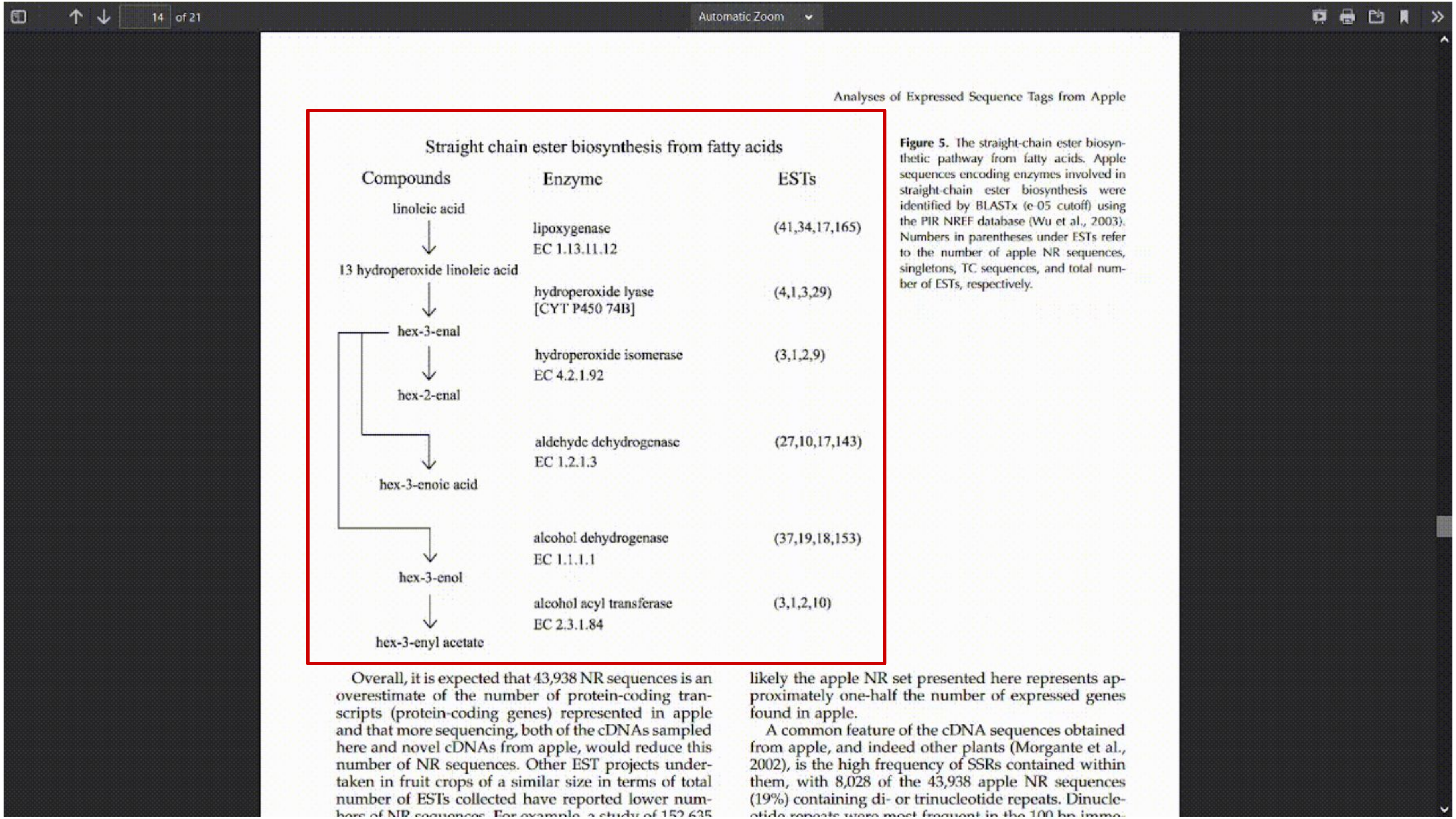


proceeds to leucoanthocyanidins produced by dihydroflavonol reductase (EC 1.1.1.219; eight NR sequences) then to anthocyanidins by anthocyanidin synthase (EC 1.14.11.19; six NR sequences). Finally, the red-colored cyanidin 3-glycosides are formed through the transfer of a sugar onto a hydroxyl group by a glycosyl transferase (EC 2.4.1.91; 26 NR sequences). Also, from dihydroquercetin, quercetin can be synthesized by flavonol synthase (EC 1.14.11.23; seven NR sequences), which in turn can be glycosylated by glycosyl transferases. Some members of these gene families that are expressed in the apple skin have been isolated and shown to be inducible by UV and coordinately up-regulated in the skins of red apple varieties (Kim et al., 2003; Ben-Yehudah et al., 2005). In addition, members of the MYB transcription factor family have been identified that can interact with promoters of these genes (Hellens et al., 2005). Furthermore, one MYB (MdMYB10; one NR sequence) has been identified that up-regulates this pathway in apple skin (R.V. Espley, R.P. Hellens, J. Putterill, and A.C. Allan, personal communication).

Gene Family Evolution

Within the apple sequence dataset, there are representatives of many large gene families involved in the biosynthesis of phytochemicals, such as the flavor and health compounds described above. Such multigene families include the acyl transferases, methyl transferases, glycosyl transferases, and cytochrome P450s. We have compared the predicted amino acid sequences of members of selected biosynthetic gene families from Arabidopsis and apple using phylogenetic methods to identify clades where apple genes may have expanded in number, presumably by gene duplication. An example of this type of analysis is shown for the acyl transferases (Fig. 8), a gene family that contains members that are involved in ester biosynthesis in apple (Souleyre et al., 2005). For this gene family, there is at least one clade with more representatives from apple than Arabidopsis. Expansions of gene number in apple are frequent in genes that are found in fruit cDNA libraries. For example, there have been more expansions of clades in apple P450s that contain

Figure 3. The sorbitol metabolism pathway in apple. Apple sequences encoding enzymes involved in sorbitol metabolism were identified by BLASTx (e-05 cutoff) using the PIR NREF database (Wu et al.,



Straight chain ester biosynthesis from fatty acids

Compounds	Enzyme	ESTs
linoleic acid		
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		

Figure 5. The straight-chain ester biosynthetic pathway from fatty acids. Apple sequences encoding enzymes involved in straight-chain ester biosynthesis were identified by BLASTx (e-05 cutoff) using the PIR NREF database (Wu et al., 2003). Numbers in parentheses under ESTs refer to the number of apple NR sequences, singletons, TC sequences, and total number of ESTs, respectively.

Overall, it is expected that 43,938 NR sequences is an overestimate of the number of protein-coding transcripts (protein-coding genes) represented in apple and that more sequencing, both of the cDNAs sampled here and novel cDNAs from apple, would reduce this number of NR sequences. Other EST projects undertaken in fruit crops of a similar size in terms of total number of ESTs collected have reported lower numbers of NR sequences. For example, a study of 152,635 likely the apple NR set presented here represents approximately one-half the number of expressed genes found in apple.

A common feature of the cDNA sequences obtained from apple, and indeed other plants (Morgante et al., 2002), is the high frequency of SSRs contained within them, with 8,028 of the 43,938 apple NR sequences (19%) containing di- or trinucleotide repeats. Dinucleotide repeats were most frequent in the 100 bp imme-

Goal

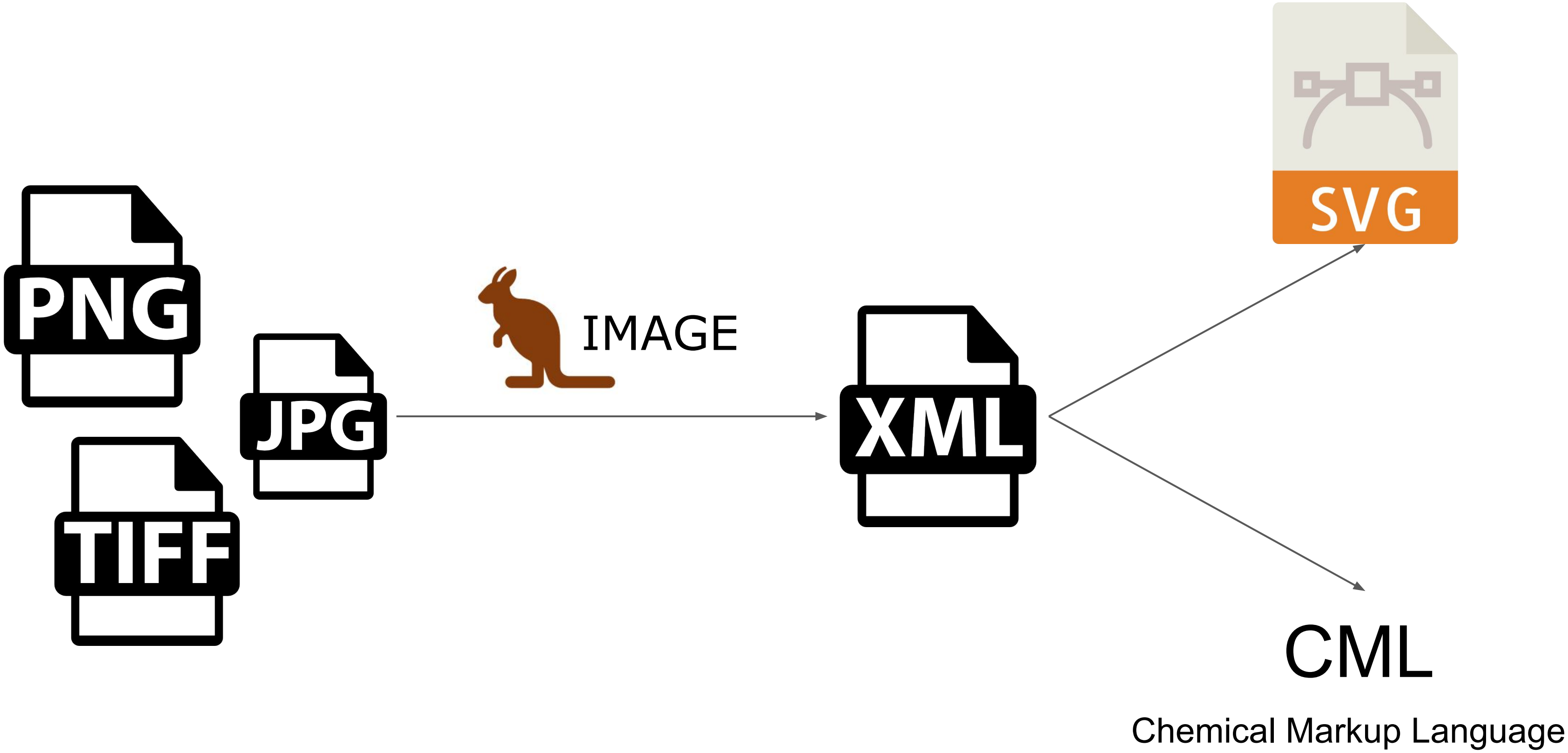
Useful scientific information is often locked up in images in scientific papers

Currently you need a human to read and interpret the diagrams

IMAGE

Our goal with **pyamiimage** is to read such images and extract useful semantic information from the images **automatically**!

Right now we are working on **biosynthetic pathways**



IMAGE

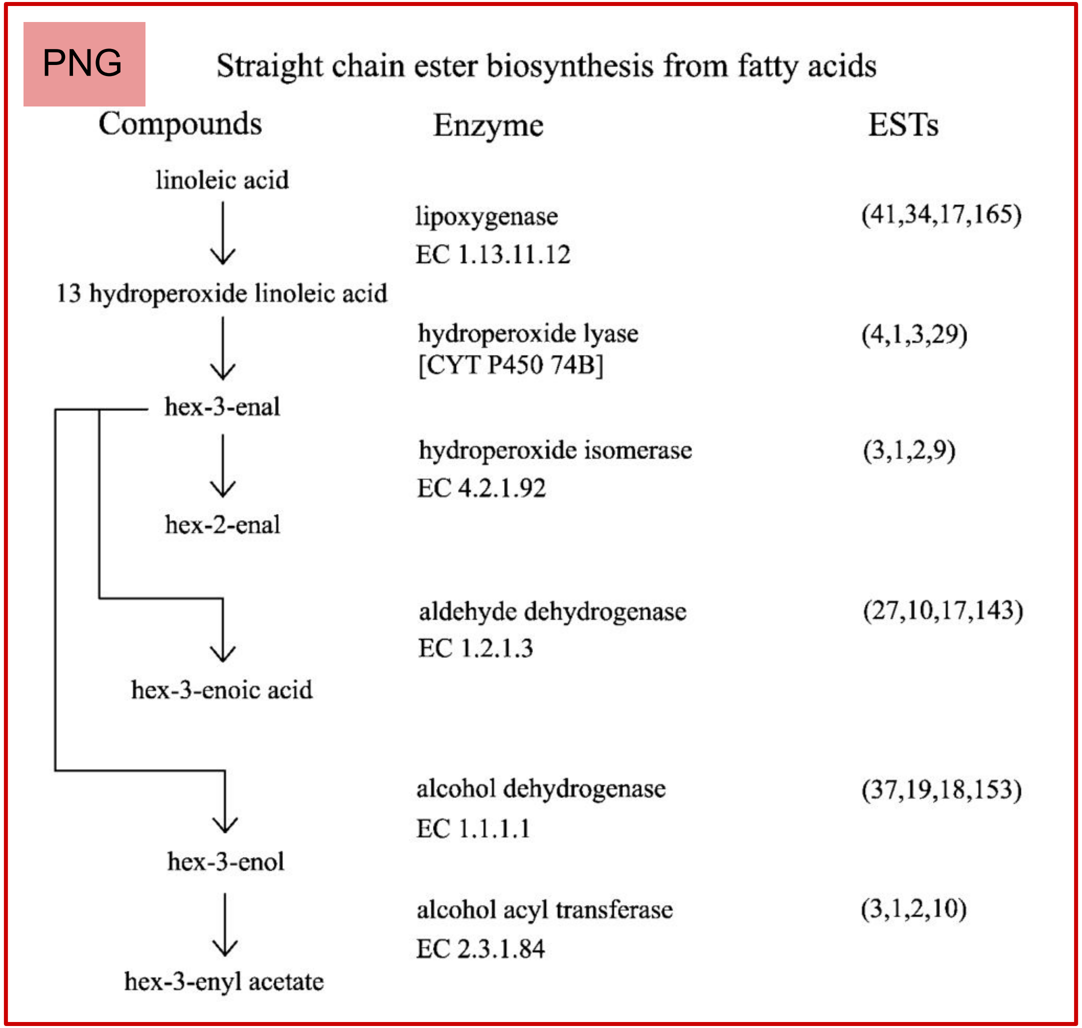
CML

Chemical Markup Language

A biosynthetic pathway image found in literature

This is a DUMB document Just pixels, only humans can read and understand it

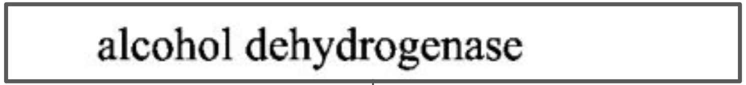
Let’s try to make it smarter Extract text with Tesseract



PNG

Straight chain ester biosynthesis from fatty acids

Compounds	Enzyme	ESTs
linoleic acid		
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		

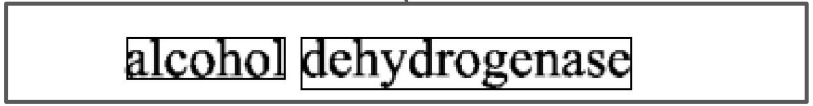


alcohol dehydrogenase

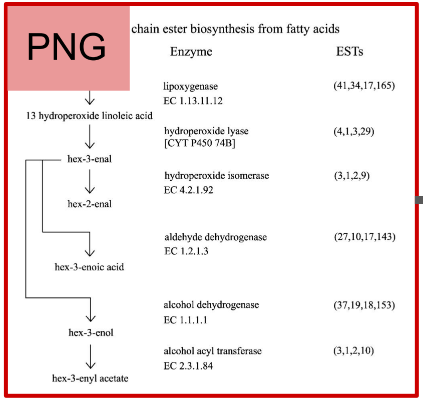
Tesseract OCR

tesseract-ocr / tesseract

```
<span class="ocr_line" id="line_1_25" title="bbox 392 738 629 762; baseline 0 -5; x_size 24; x_descenders 5; x_ascenders 6">
<span class="ocrx_word" id="word_1_45" title="bbox 392 738 466 757; x_wconf 96">alcohol</span>
<span class="ocrx_word" id="word_1_46" title="bbox 474 738 629 762; x_wconf 96">dehydrogenase</span>
</span>
```



alcohol dehydrogenase



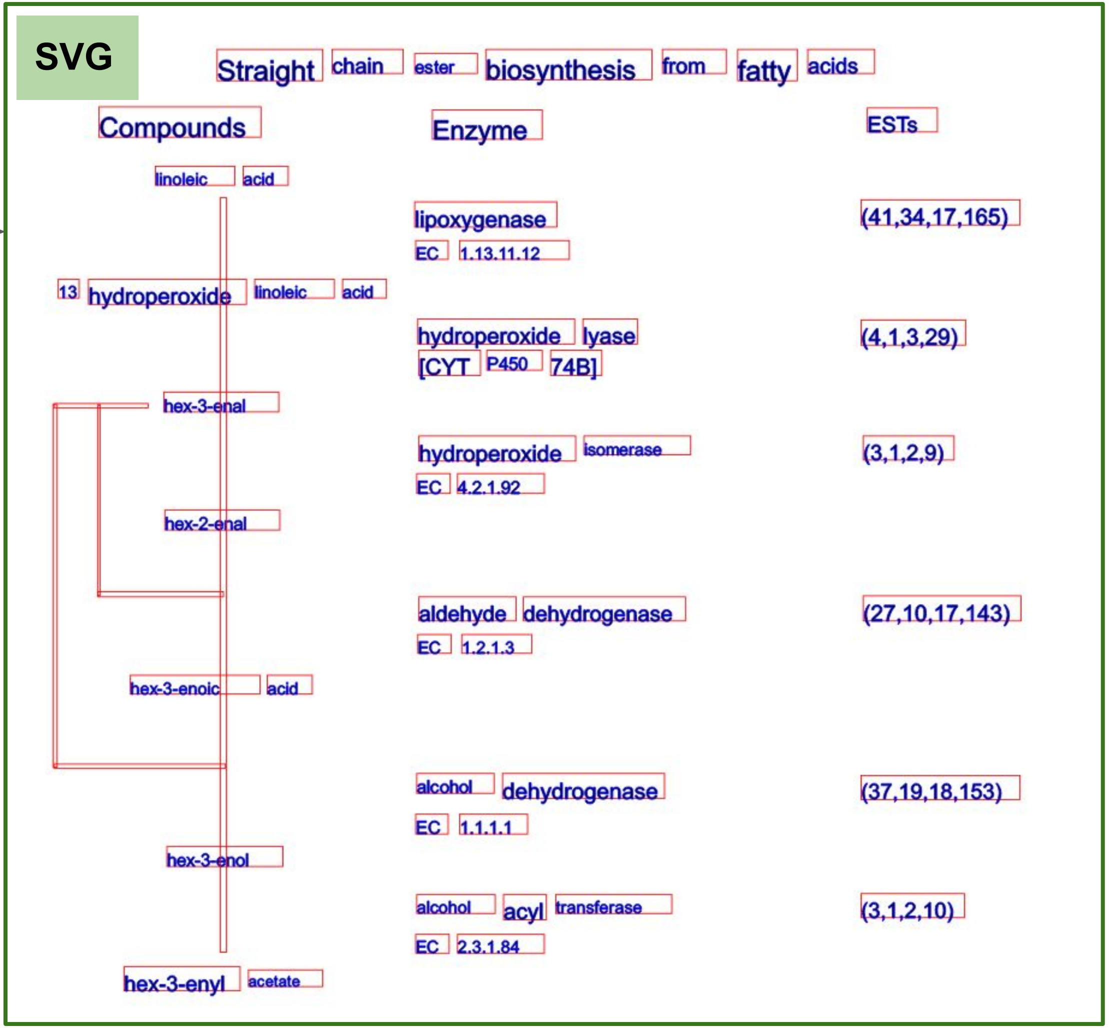
PNG

chain ester biosynthesis from fatty acids

	Enzyme	ESTs
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		

SVG from Tesseract output

Tesseract recognises text in the whole image

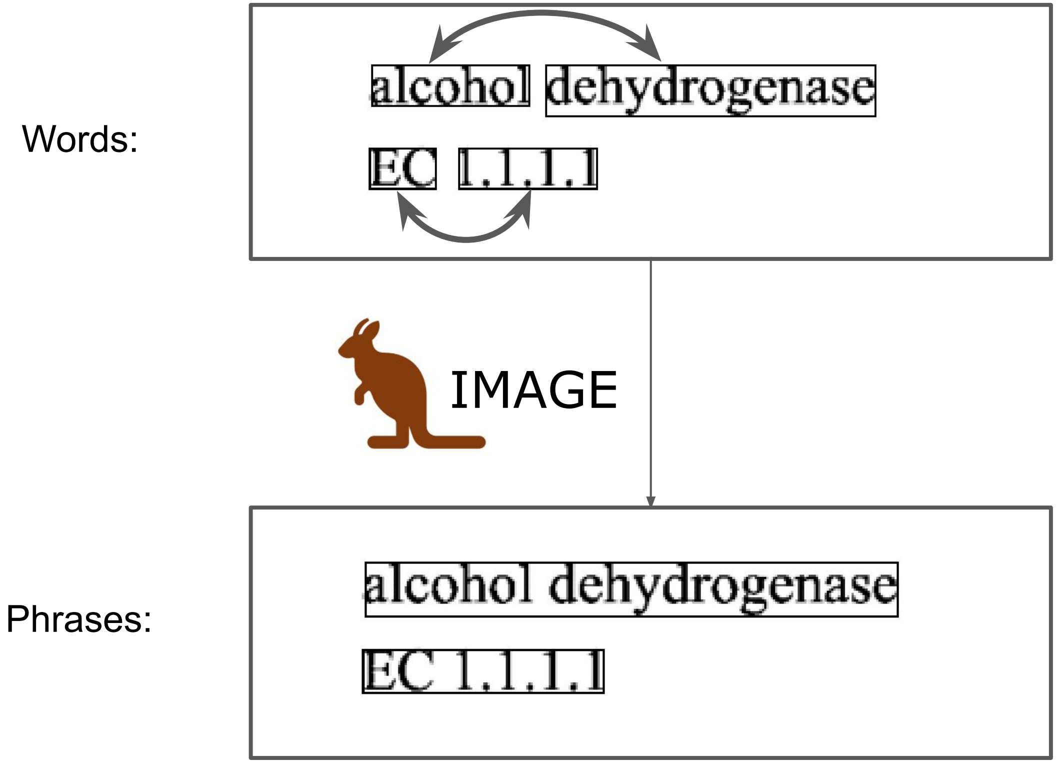


SVG

Straight chain ester biosynthesis from fatty acids

Compounds	Enzyme	ESTs
linoleic acid		
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		

Geometric Phrase Detection

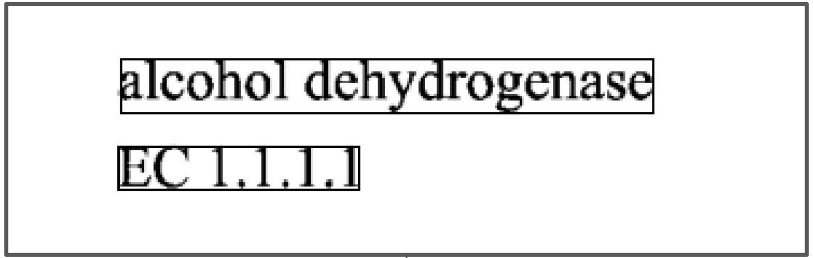


Words:

IMAGE

Phrases:

Wikidata/KEGG/PubChem Lookup



alcohol dehydrogenase

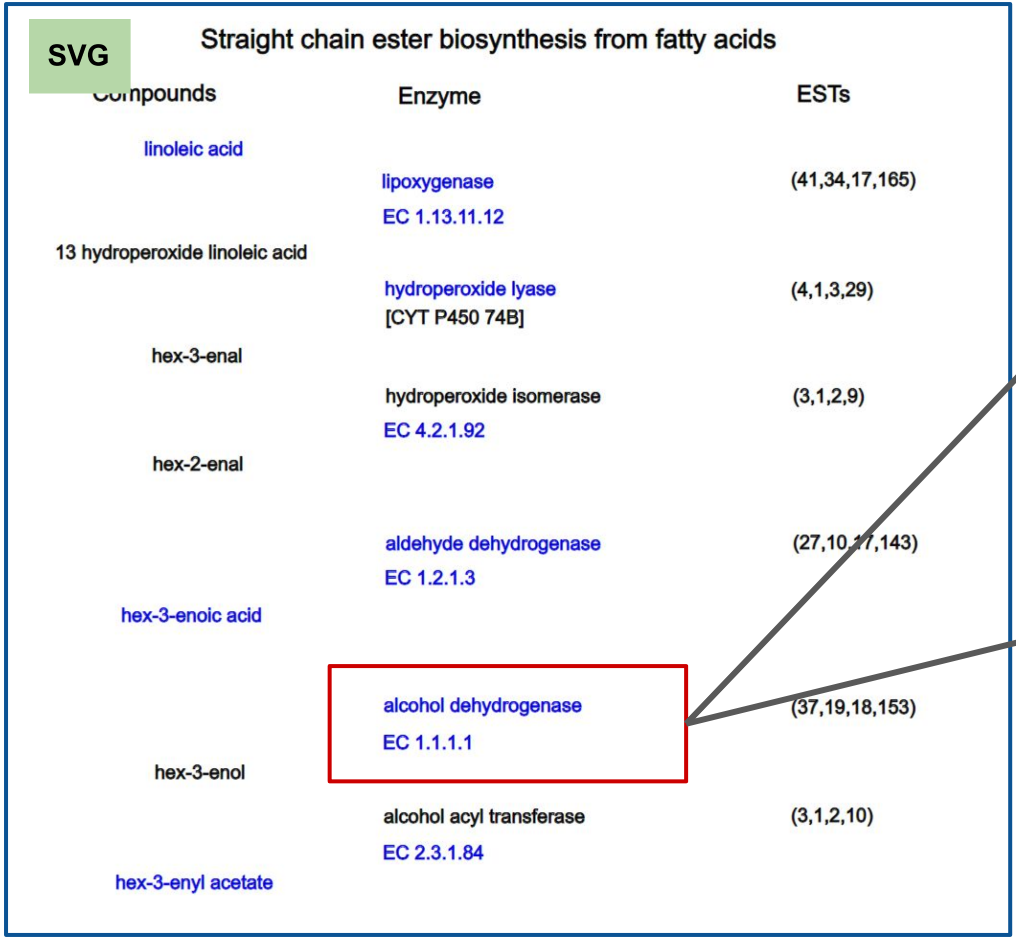
EC 1.1.1.1

IMAGE

Wikidata
https://www.wikidata.org/wiki/Q410754

KEGG
https://www.genome.jp/entry/1.1.1.1

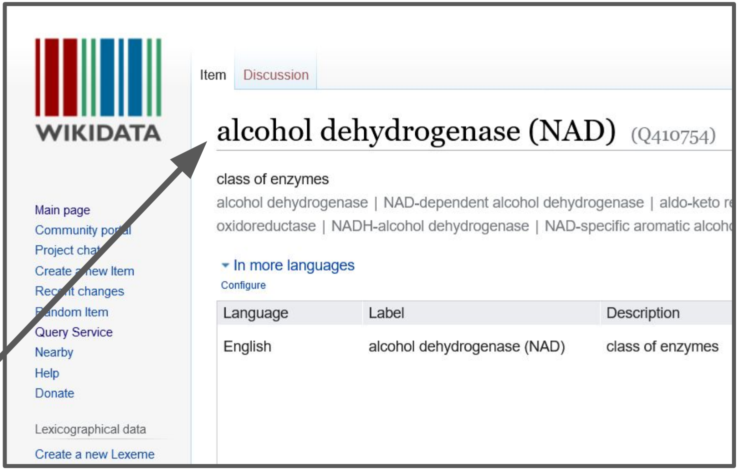
Automatic hyperlinks



SVG

Straight chain ester biosynthesis from fatty acids

Compounds	Enzyme	ESTs
linoleic acid		
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		



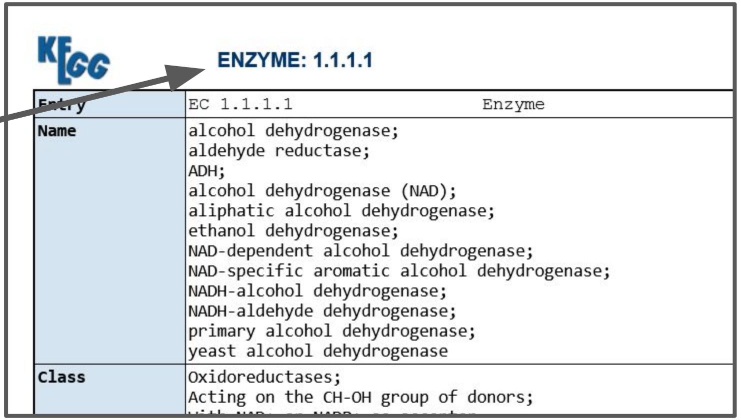
alcohol dehydrogenase (NAD) (Q410754)

class of enzymes

alcohol dehydrogenase | NAD-dependent alcohol dehydrogenase | aldo-keto re
oxidoreductase | NADH-alcohol dehydrogenase | NAD-specific aromatic alcoho

▾ In more languages
Configure

Language	Label	Description
English	alcohol dehydrogenase (NAD)	class of enzymes



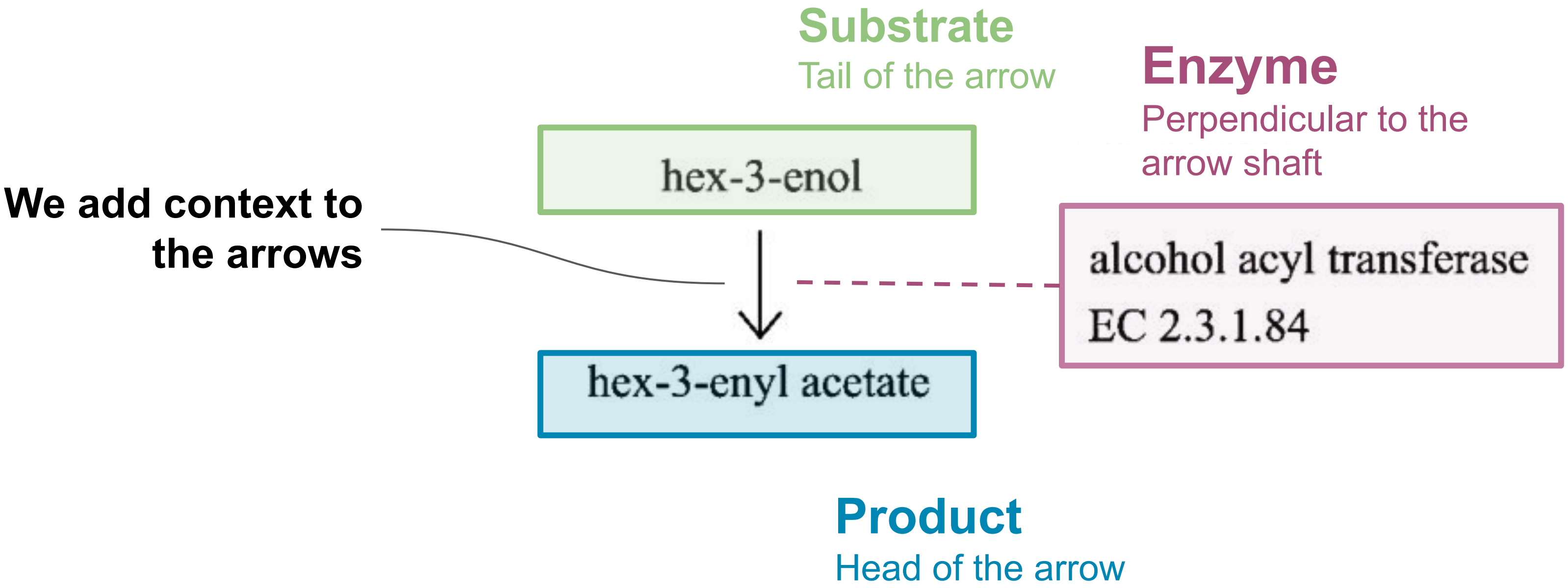
ENZYME: 1.1.1.1

Entry	EC 1.1.1.1 Enzyme

Name	alcohol dehydrogenase;
	aldehyde reductase;
	ADH;
	alcohol dehydrogenase (NAD);
	aliphatic alcohol dehydrogenase;
	ethanol dehydrogenase;
	NAD-dependent alcohol dehydrogenase;
	NAD-specific aromatic alcohol dehydrogenase;
	NADH-alcohol dehydrogenase;
	NADH-aldehyde dehydrogenase;
	primary alcohol dehydrogenase;
	yeast alcohol dehydrogenase

Class	Oxidoreductases;
	Acting on the CH-OH group of donors;

Arrows in Pathways



Substrate

Tail of the arrow

Enzyme

Perpendicular to the arrow shaft

hex-3-enol

We add context to the arrows

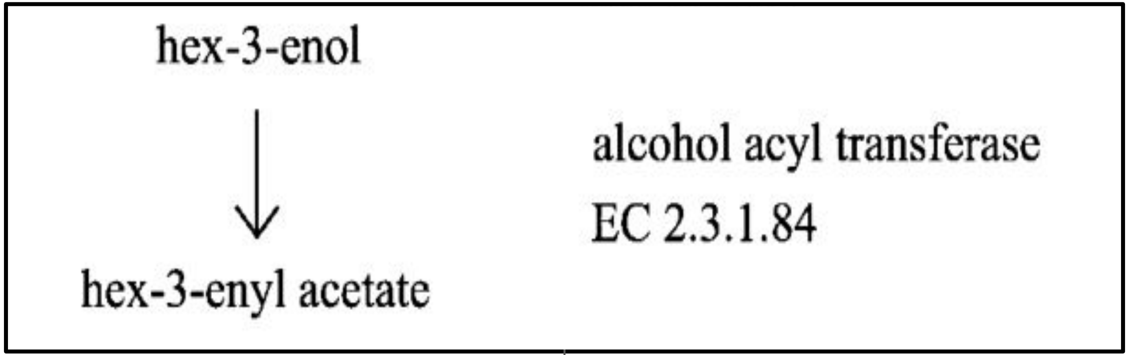
alcohol acyl transferase

EC 2.3.1.84

hex-3-enyl acetate

Product

Head of the arrow

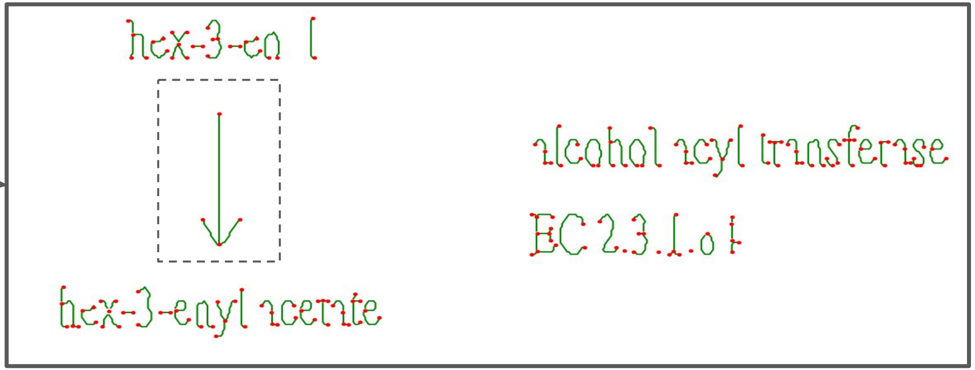


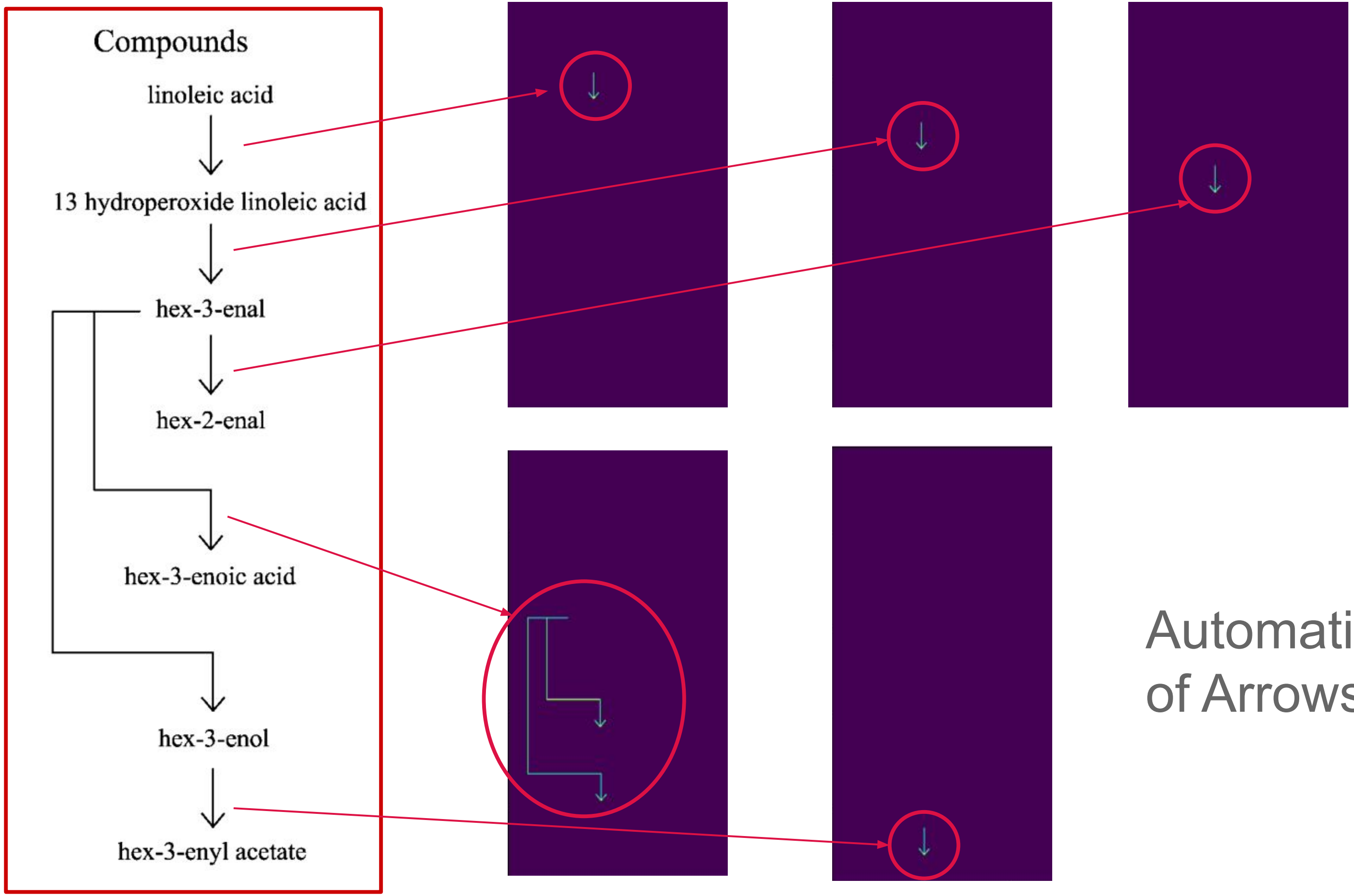
Invert ~> Binarize ~> Skeletonize

skimage + **Image-Py / sknw**

IMAGE

Geometric
Arrow detection





Compounds

linoleic acid

13 hydroperoxide linoleic acid

hex-3-enal

hex-2-enal

hex-3-enoic acid

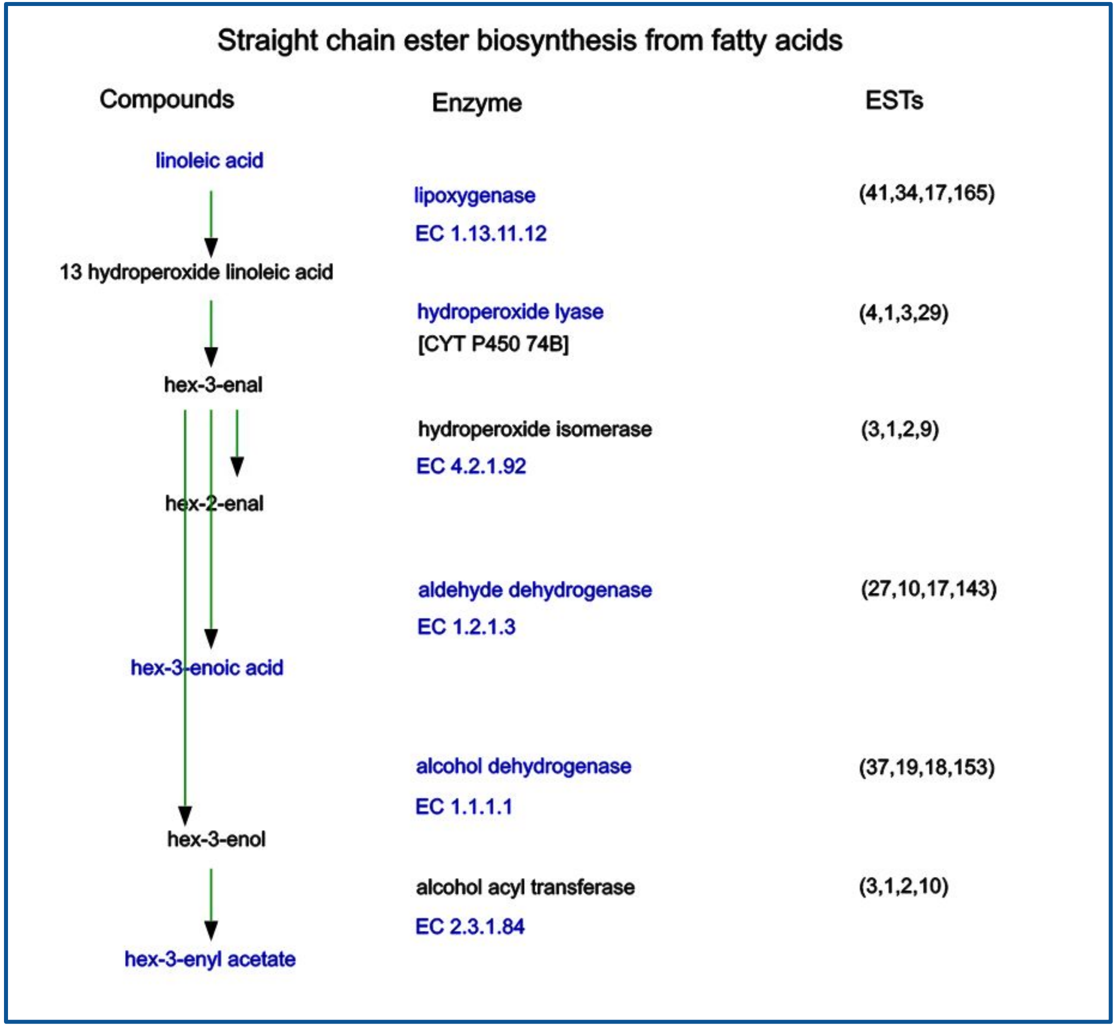
hex-3-enol

hex-3-enyl acetate

Automatic detection of Arrows

SVG with links embedded in the text and arrows

Conversion to **CMLReact** will be automatic



Straight chain ester biosynthesis from fatty acids

Compounds	Enzyme	ESTs
linoleic acid		
	lipoxygenase EC 1.13.11.12	(41,34,17,165)
13 hydroperoxide linoleic acid		
	hydroperoxide lyase [CYT P450 74B]	(4,1,3,29)
hex-3-enal		
	hydroperoxide isomerase EC 4.2.1.92	(3,1,2,9)
hex-2-enal		
	aldehyde dehydrogenase EC 1.2.1.3	(27,10,17,143)
hex-3-enoic acid		
	alcohol dehydrogenase EC 1.1.1.1	(37,19,18,153)
hex-3-enol		
	alcohol acyl transferase EC 2.3.1.84	(3,1,2,10)
hex-3-enyl acetate		

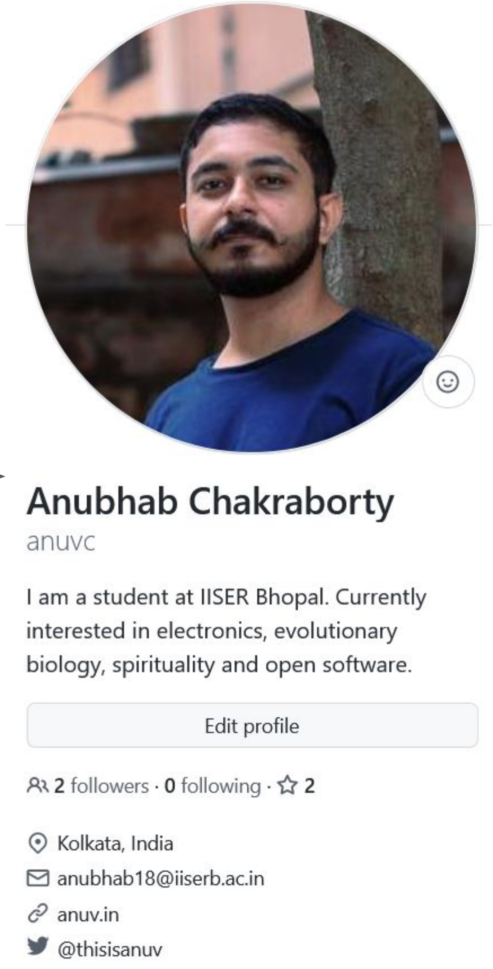
Thank you for listening!

Find us in Github

The code for IMAGE can be found in:

https://github.com/petermr/pyamiimage

Github: **@anuvc**



Anubhab Chakraborty

anuvc

I am a student at IISER Bhopal. Currently interested in electronics, evolutionary biology, spirituality and open software.

Edit profile

2 followers · 0 following · ☆ 2

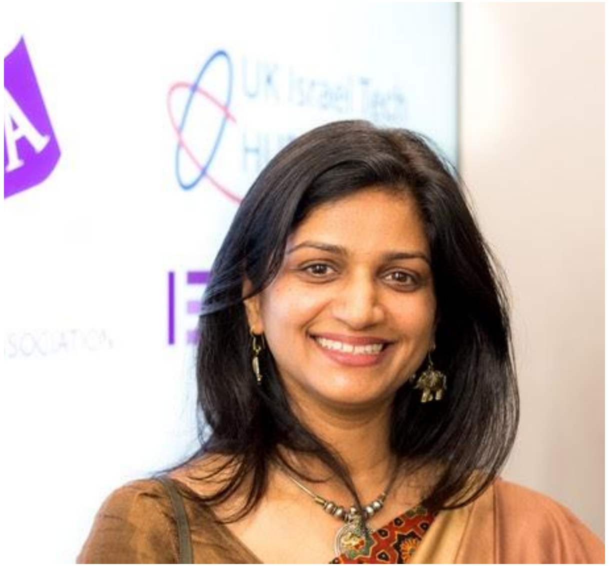
Kolkata, India

anubhab18@iiserb.ac.in

anuv.in

@thisisanuv

Thanks to my mentors!



Dr. Peter Murray-Rust
Github: **@petermr**

Dr. Gitanjali Yadav
Github: **@gilienv**